UNIVERSIDADE FEDERAL DA FRONTEIRA SUL
CAMPUS CHAPECÓ
CURSO DE CIÊNCIA DA COMPUTAÇÃO

JOÃO MARCOS CAMPAGNOLO

AVALIAÇÃO DA SENSIBILIDADE DE MÉTRICAS DE AVALIAÇÃO DE TÓPICOS

CHAPECÓ
2018

JOÃO MARCOS CAMPAGNOLO

AVALIAÇÃO DA SENSIBILIDADE DE MÉTRICAS DE AVALIAÇÃO DE TÓPICOS

Trabalho de conclusão de curso de graduação apresentado como requisito para obtenção do grau de Bacharel em Ciência da Computação da Universidade Federal da Fronteira Sul.
Orientador: Prof. Dr. Denio Duarte

CHAPECÓ
2018

Bibliotecas da Universidade Federal da Fronteira Sul - UFFS

Campagnolo, João Marcos
Avaliação da Sensibilidade de Métricas de Avaliação de Tópicos / João Marcos Campagnolo. -- 2018.
55 f.:il.

Orientador: Dr. Denio Duarte.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal da Fronteira Sul, Curso de Ciência da Computação, Chapecó, SC , 2018.

1. Modelagem de Tópicos. 2. Métrica de Coerência. 3. Tópicos. 4. Métricas. 5. Sensibilidade. I. Duarte, Denio, orient. II. Universidade Federal da Fronteira Sul. III. Título.

Elaborada pelo sistema de Geração Automática de Ficha de Identificação da Obra pela UFFS com os dados fornecidos pelo(a) autor(a).

JOÃO MARCOS CAMPAGNOLO

AVALIAÇÃO DA SENSIBILIDADE DE MÉTRICAS DE AVALIAÇÃO DE TÓPICOS

Trabalho de conclusão de curso de graduação apresentado como requisito para obtenção do grau de Bacharel em Ciência da Computação da Universidade Federal da Fronteira Sul.

Orientador: Prof. Dr. Denio Duarte

Aprovado em: 07\12\2018

BANCA EXAMINADORA:

Dr. Denio Duarte - UFFS

Dr. Guilherme Dal Bianco - UFFS

Ma. Andressa Sebben - UFFS

RESUMO

A crescente tendência de armazenar todo o conhecimento e conteúdo produzido de forma digital dificulta cada vez mais a tarefa de buscar e organizar as informações. Os algoritmos de modelagem de tópicos permitem extrair temas/assuntos, em forma de tópicos, de vastas coleções de documentos. Um documento, que pode ser curto ou longo, pode ser definido como uma mistura de tópicos, que são um conjunto de palavras ordenadas por suas probabilidades de ocorrência. Avaliar a qualidade de um tópico é uma tarefa que pode ser simples para seres humanos, tornando-se muito custosa em se tratando de grandes quantidades de dados. Dessa forma, métodos computacionais, conhecidos como métricas de coerência, são utilizados para medir a qualidade de tópicos a partir da co-ocorrência entre as palavras que os compõem. Porém, diferentes métricas podem gerar diferentes resultados quando aplicadas a um mesmo tópico. Neste trabalho será realizada uma avaliação da sensibilidade de algumas dessas métricas, aplicando-as em um conjunto de tópicos que foram criados, deturpados através da inserção de palavras intrusas, e validados por seres humanos. Como resultado, de modo geral, a métrica C_{UCI} se mostrou ser a mais sensível, enquanto as métricas C_V e C_{UMASS} se mostraram as menos sensíveis.

Palavras-chave: Modelagem de Tópicos; Tópicos; Métricas de Coerência; Métricas; Sensibilidade.

ABSTRACT

The growing tendency of store all the knowledge and content produced digitally makes it increasingly difficult to find all this information and organize it. Topical modeling algorithms allows to extract topics from vast collections of documents. A document, which may be short or long, can be defined as a mixture of topics, which are a set of words sorted by their probability of occurrence. Evaluating the quality of a topic is a task that can be simple for humans, although it is very expensive when dealing with large amounts of data. Thus, computational methods, known as coherence metrics, are used to measure the quality of topics from the co-occurrence between the words that compose them. However, different metrics can generate different results when applied to the same topic. This work will evaluate the sensitivity of some of these metrics by applying them to a set of topics that have been created, adulterated by inserting intrusive words, and validated by humans. As result, in general, the metric C_{UICI} was shown to be the most sensitive, while the metrics C_V and C_{UMASS} were shown to be the least sensitive.

Keywords: Topic Modeling. Topics. Coherence Measures. Metrics. Sensitivity.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

LDA	*Latent Dirichlet Allocation*
LSA	*Latent Semantic Analysis*
NPMI	*Normalized Pointwise Mutual Information*
PLSA	*Probabilistic Latent Semantic Analysis*
PMI	*Pointwise Mutual Information*

SUMÁRIO

1 INTRODUÇÃO

1.1 Tema

A modelagem de tópicos consiste na varredura de uma coleção de documentos, visando extrair um conjunto de palavras que descrevem um tema ou um assunto. No contexto deste trabalho, o conjunto de palavras que descreve um tema/assunto é chamado de tópico.

Algoritmos de Modelagem de Tópicos são classificados como não-supervisionados (sem rótulo), tornando difícil avaliar a qualidade dos tópicos extraídos pelos modelos. Medir a coerência dos tópicos gerados é um problema que vem sendo muito estudado recentemente, uma vez que não há garantia na interpretação correta dos tópicos extraídos. (RÖDER; BOTH; HINNEBURG, 2015).

O tema deste trabalho é identificar e quantificar a sensibilidade de algumas das métricas propostas por RÖDER; BOTH; HINNEBURG (2015), afim de conhecer qual métrica é mais ou menos sensível em cada cenário proposto.

1.2 Delimitação do Problema

Cada vez mais o conhecimento coletivo vem sendo armazenado digitalmente, na forma de notícias, blogues, artigos científicos, livros, imagens, sons, vídeos e redes sociais, o que torna cada vez mais difícil encontrar o conteúdo que se deseja (BLEI, 2012). Daí a necessidade de se criar novas ferramentas que ajudem a organizar essa vasta quantidade de conteúdo. A extração de tópicos, que podem ser interpretados como assuntos recorrentes em documentos, é uma abordagem que pode ser utilizada para sanar este problema. Os algoritmos de modelagem de tópicos são uma solução para tal, porém, a natureza não-supervisionada desses algoritmos torna difícil a avaliação da corretude interpretativa dos tópicos gerados, tornando dificultosa a tarefa de avaliar a qualidade dos mesmos.

Segundo BLEI (2012), avaliar a qualidade de um tópico manualmente não é viável, uma vez que não temos o poder humano necessário para ler e estudar vastas quantidades de informações. O uso de ferramentas computacionais que implementam métricas de coerência para avaliar a qualidade de um tópico é uma das alternativas para tal. Porém, essas métricas são capazes apenas de medir a similaridade das palavras contidas em um determinado tópico através da co-ocorrência das mesmas. O problema disso é que diferentes métricas podem tratar as co-

ocorrências de formas diferentes, uma vez que utilizam medidas diferentes de similaridade entre palavras, gerando resultados divergentes umas das outras.

1.3 Justificativa

Embora os algoritmos de modelagem de tópicos sejam capazes de extrair tópicos de conjuntos de documentos, avaliar a qualidade dos mesmos é uma tarefa difícil de se realizar computacionalmente. Segundo RÖDER; BOTH; HINNEBURG (2015), medir a coerência dos tópicos gerados é um problema que vem sendo estudado recentemente, uma vez que os modelos não garantem a interpretabilidade correta das saídas encontradas.

A divergência nos resultados obtidos mediante a utilização de diversas métricas, para os mesmos tópicos, se dá pelo fato de que cada uma mede a similaridade das palavras de forma diferente. Enquanto para algumas métricas a presença de palavras intrusas (palavras com probabilidade significativamente mais baixa que as demais) pode resultar em um nota muito baixa, para outras, essas palavras podem não causar tanto impacto.

É nesse contexto que avaliar a sensibilidade das métricas de coerência se mostra pertinente. Estudar o comportamento dessas métricas e identificar quais delas são mais ou menos sensíveis em determinadas situações, pode auxiliar na escolha da métrica mais adequada a ser utilizada.

1.4 Objetivos

1.4.1 Objetivo Geral

O objetivo deste trabalho é realizar experimentos controlados, através de diferentes cenários (configurações) de tópicos, a fim de avaliar a sensibilidade das principais métricas de avaliação de tópicos gerados por modelo probabilístico propostas por RÖDER; BOTH; HINNEBURG (2015).

1.4.2 Objetivos Específicos

- Definir os assuntos utilizados para a criação dos tópicos;
- Definir as palavras que compõe os tópicos de cada assunto;

- Criar as diferentes configurações de tópicos a partir da inserção de palavras intrusas;
- Realizar a validação dos tópicos criados;
- Definir as métricas a serem avaliadas;
- Realizar a aplicação das métricas de forma controlada;
- Definir uma métrica que seja capaz de quantificar a sensibilidade;
- Analisar e apresentar os resultados obtidos.

1.5 Estrutura do Trabalho

A sequência deste trabalho está estruturada da seguinte forma: o Capítulo 2 discorre sobre *Modelagem de Tópicos*, apresentando conceitos sobre *Documentos* e *Tópicos*. No Capítulo 3 são apresentadas as *Métricas* e a ferramenta *Palmetto* que foram propostas por RÖDER; BOTH; HINNEBURG (2015). O Capítulo 4 apresenta o projeto de experimento que define as etapas e procedimentos que foram utilizados para atingir os objetivos definidos. O Capítulo 5 apresenta as análises e resultados obtidos. Por fim, no Capítulo 6 são apresentadas as conclusões.

2 MODELAGEM DE TÓPICOS

Algoritmos de Modelagem de Tópicos são métodos que analisam as palavras contidas em um conjunto de documentos a fim de descobrir os temas que estão associados aos mesmos. Tais algoritmos não necessitam de entradas rotuladas (são algoritmos de aprendizagem não-supervisionada), uma vez que os tópicos são gerados a partir dos textos originais dos documentos. Em (STEYVERS; GRIFFITHS, 2007), é definido um modelo de tópicos como sendo um modelo generativo para documentos. Tal modelo especifica um procedimento probabilístico simples pelo qual os documentos podem ser gerados. Os tópicos, por sua vez, são compostos por uma distribuição probabilística de palavras, sendo que documentos com conteúdos diferentes podem ser gerados a partir de distribuições diferentes sobre os tópicos. Ainda, é possível inferir o conjunto de tópicos responsáveis por gerar uma coleção de documentos através de técnicas estatísticas que fazem o processo inverso.

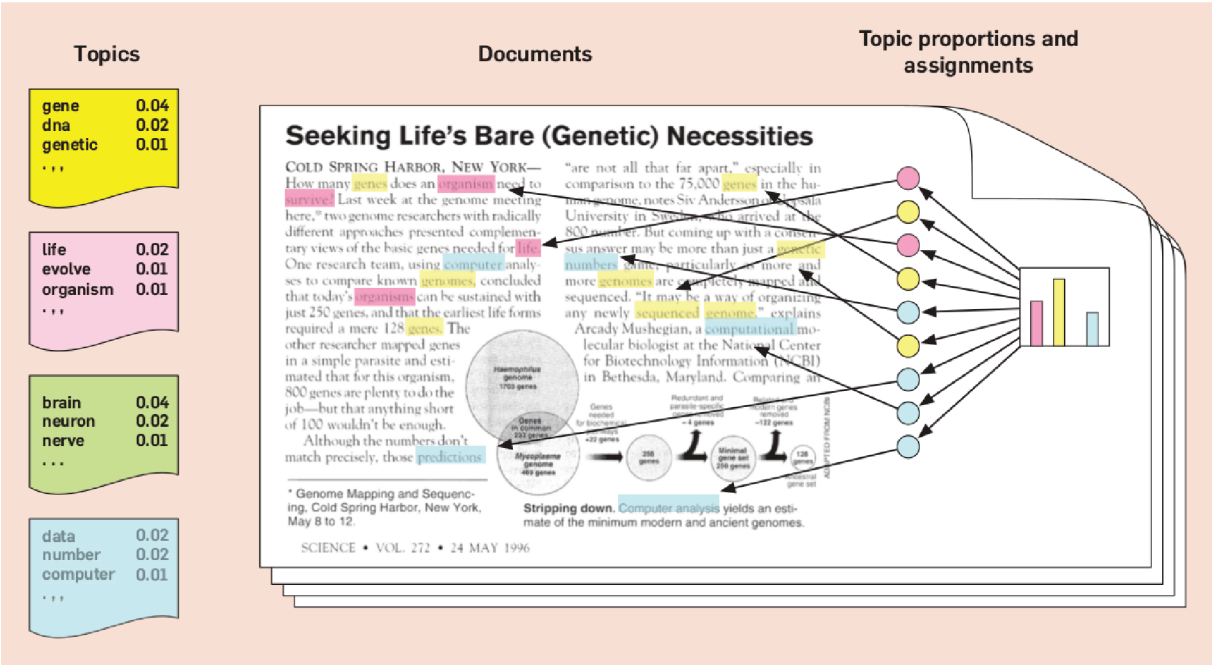

Figura 2.1 – Exemplo de Modelagem de Tópicos (BLEI, 2012).

O exemplo da Figura 2.1 mostra um artigo intitulado *Seeking Life's Bare (Genetic) Necessities*, que discorre sobre o uso da análise de dados na determinação do número de genes que um organismo precisa para sobreviver. Manualmente, palavras sobre análise de dados foram destacadas em azul, palavras sobre biologia evolucionária foram destacadas em rosa, e palavras sobre genética foram destacadas em amarelo. Segundo BLEI (2012), se destacadas todas as

palavras do artigo (desconsiderando as *stop-words* - palavras irrelevantes à extração de tópicos, tais como preposições e artigos), seria possível inferir que o artigo combina os tópicos de genética, biologia evolucionária e análise de dados.

A Modelagem de Tópicos, através da extração de tópicos, viabiliza a organização e sumarização de arquivos digitais em um escala impossível aos seres humanos (BLEI, 2012), além de poder ser utilizada em vastos conjuntos de documentos, podendo ser adaptada para diferentes tipos de dados, tais como: dados genéticos, imagens e redes sociais.

2.1 Tópicos

Tópicos são conjuntos de palavras que ocorrem frequentemente em documentos que estão semanticamente relacionados entre si, de forma que façam sentido dentro de um contexto específico. Segundo BLEI (2012), pode-se definir um tópico formalmente como sendo a distribuição sobre um vocabulário fixo. Por exemplo, o tópico *genética* possui palavras relacionadas à genética com uma maior probabilidade de acontecerem, enquanto o tópico *biologia evolucionária* possui palavras relacionadas à biologia evolucionária com uma maior probabilidade.

"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel

Figura 2.2 – Exemplo de tópicos gerados (BLEI, 2012).

Na Figura 2.2 estão representados quatro tópicos distintos, cada um com as *top-10 words* (dez palavras com maior probabilidade de ocorrência dentro daquele tópico), sendo que as palavras estão ordenadas de acordo com a sua probabilidade de ocorrência. Tomando como exemplo o tópico *Computers*, as palavras *computer*, *models*, *information* e *data*, nesta ordem, são as palavras com maior probabilidade de ocorrência no referido tópico.

2.2 Documentos

BLEI (2012) define um documento como sendo uma mistura de tópicos, onde um tópico é, por sua vez, uma distribuição de probabilidade sobre palavras. Um dado documento pode estar relacionado com um ou mais tópicos, e pode ser considerado curto ou longo, dependendo de sua extensão. Por exemplo, uma postagem em rede social pode ser considerada um documento curto, já um artigo científico em uma revista pode ser considerado um documento longo.

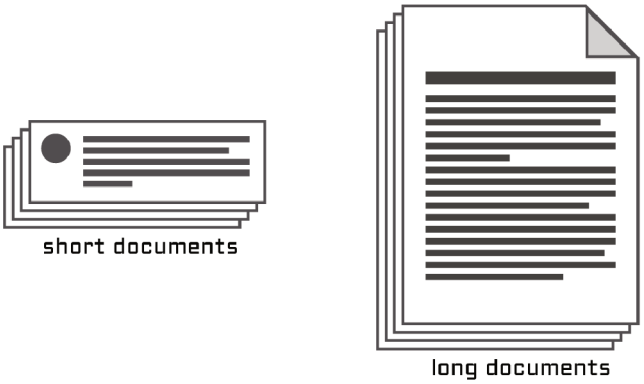

Figura 2.3 – Exemplo de documento curto e documento longo.

Antes de passar pela modelagem de tópicos, cada documento original precisa passar por algumas etapas de pré-processamento, resultando assim no conjunto de documentos que será de fato utilizado. Essas etapas consistem basicamente na eliminação das *stop-words* (palavras irrelevantes no contexto de extração de tópicos) e na tradução das palavras para o seu radical. Outras etapas podem ser necessárias, dependendo do tipo de dados dos documentos a serem utilizados.

Para a maioria das abordagens de modelagem de tópicos, como por exemplo, *Latent Dirichlet Allocation (LDA)*, *Latent Semantic Analysis (LSA)* e *Probabilistic Latent Semantic Analysis (PLSA)*, um documento pode ser visto como uma *bag-of-words* (sacola de palavras) onde a ordem em que as palavras aparecem não importa.

3 MÉTRICAS

Métricas são ferramentas que podem ser utilizadas para medir e avaliar resultados de experimentos. Na modelagem de tópicos, os modelos gerados são difíceis de se avaliar. Isso ocorre, pois os algoritmos de aprendizagem não-supervisionada não possuem rótulos que permitam uma verificação da corretude dos resultados obtidos (CHANG et al., 2009).

Segundo RÖDER; BOTH; HINNEBURG (2015), o interesse em métricas capazes de avaliar a coerência dos resultados obtidos na extração de tópicos, a partir de algoritmos de aprendizagem não-supervisionada, surge do fato de que esses algoritmos de modelagem de tópicos não garantem a interpretabilidade correta dos modelos gerados. Além disso, a avaliação dos tópicos realizada por meio de seres humanos se torna um meio muito caro e custoso para se reproduzir (BLEI, 2012). É neste contexto que RÖDER; BOTH; HINNEBURG (2015) propõem uma ferramenta que implementa algumas métricas já conhecidas e que buscam ser coerentes com bases de dados de ranqueamento gerados por seres humanos.

A seguir serão apresentados alguns conceitos, métodos e as métricas propostas por RÖDER; BOTH; HINNEBURG (2015). Para fins de exemplificação, o seguinte conjunto de palavras $t_1 = \{car, driver, wheel, speed\}$, representando um tópico ordenado pela probabilidade de ocorrência das palavras, será utilizado no decorrer deste capítulo.

3.1 *Sliding Window*

Uma *sliding window* (ou janela deslizante) é um subconjunto de palavras consecutivas, de tamanho N, que pode ser deslocado palavra por palavra para qualquer um dos lados. Como exemplo, a Figura 3.1 mostra 3 possibilidades diferentes de uma *sliding window* de tamanho 4:

$$doc_1 = \{control \underbrace{drive}_{sw_y} \underbrace{car\ speed\ park}_{sw_g} passenger \underbrace{comfort\ safety\ wheel\ crash}_{sw_r}\}$$

Figura 3.1 – Exemplo de *sliding window* de tamanho 4.

Na Figura 3.1, a *sliding window* destacada em amarelo (sw_y) é o subconjunto contendo as palavras *drive*, *car*, *speed* e *park*. Já a *sliding window* destacada em verde (sw_g) é o subconjunto contendo as palavras *car*, *speed*, *park* e *passenger*. Podemos dizer que as palavras *speed* e *park* estão contidas em ambas *sliding windows* (amarela e verde). A *sliding window* destacada

em vermelho (sw_r) não compartilha nenhuma palavra com as outras duas *sliding windows*.

3.2 *Context Window*

Uma *context window* (ou janela de contexto) é um subconjunto com as N palavras consecutivas localizadas imediatamente ao lado de uma determinada palavra. A Figura 3.2 mostra a *context window* de tamanho 3 da palavra *park*:

$$doc_1 = \{control \underbrace{drive\ car\ speed}_{\text{CW}_g}\ \underbrace{park}_{\text{CW}_y}\ \underbrace{passenger\ comfort\ safety}_{\text{CW}_g}\ wheel\ crash\}$$

Figura 3.2 – Exemplo de *context window* de tamanho 3.

Na Figura 3.2, a *context window* da palavra *park*, destacada em amarelo (cw_y), é o subconjunto contendo as palavras *drive*, *car*, *speed*, *passenger*, *comfort* e *safety*, destacada em verde (cw_g).

3.3 *Pointwise Mutual Information (PMI)*

Pointwise Mutual Information (PMI) é uma abordagem utilizada para medir a associatividade entre duas palavras. Assim, a ordem das palavras não importa, uma vez que essa é uma abordagem simétrica. A PMI é calculada a partir de uma contagem da ocorrência de palavras, e é dada pela fórmula a seguir:

$$PMI(w_i, w_j) = \log\left(\frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)}\right)$$

Dada a fórmula, $P(w_i, w_j)$ é a frequência/probabilidade de se observar as palavras w_i e w_j na mesma janela (*context/sliding window*). Já $P(w_i)$ e $P(w_j)$ são, respectivamente, a frequência/probabilidade de se observar as palavras w_i e w_j de forma separada. Quanto mais próxima a frequência de co-ocorrência de duas palavras for da ocorrência das duas palavras em separado, melhor será a pontuação para o dado par de palavras. A constante ϵ pode ser utilizada para evitar a ocorrência de logaritmo de zero.

Como exemplo, supondo que no dado tópico a palavra *car* possui uma frequência de 0.001, a palavra *driver* possui uma frequência de 0.0005, e a frequência em que as duas palavras aparecem na mesma janela é de 0.0002, utilizando um $\epsilon = 0$, a $PMI(car, driver)$ será dada por:

$$PMI(car, driver) = \log\left(\frac{0.0002 + 0}{0.001 \cdot 0.0005}\right) = \log\left(\frac{0.0002}{5e-07}\right) \cong \log(400) \cong 5.99146$$

3.4 *Normalized Pointwise Mutual Information (NPMI)*

Normalized Pointwise Mutual Information (NPMI) é uma variação da *PMI* que normaliza o valor obtido para o intervalo $[-1, 1]$, onde o limite inferior -1 significa nenhuma co-ocorrência, o valor 0 significa independência entre as duas palavras, e 1 significa uma co-ocorrência completa. A fórmula da *NPMI* é dada por:

$$NPMI(w_i, w_j) = \frac{PMI(w_i, w_j)}{-\log(P(w_i, w_j))}$$

Como exemplo, utilizando o valor da *PMI* calculada para as palavras *car* e *driver* na seção anterior, a $NPMI(car, driver)$ será dada por:

$$NPMI(car, driver) = \frac{5.99146}{-\log(0.0002)} \cong 0.70345$$

3.5 *Confirmation Measure of Fitelson's Coherence*

Denotada por m_f, essa é uma função proposta por FITELSON (2003) que é utilizada para calcular o grau de relação entre duas proposições. Essa função também é usada por RÖDER; BOTH; HINNEBURG (2015) para calcular o grau de relação entre uma palavra w_i com o subconjunto de palavras $S(i)_j$, e é definida pela seguinte fórmula:

$$m_f(w_i, S(i)_j) = \frac{P(W_i|S(i)_j) - P(W_i|\neg S(i)_j)}{P(W_i|S(i)_j) + P(W_i|\neg S(i)_j)}$$

3.6 *UCI Coherence*

A métrica *UCI Coherence* é baseada em uma *sliding window* de tamanho 10 e na *PMI* de todos os pares de palavras das *N-top words* de um tópico, sendo definida pela seguinte fórmula:

$$C_{UCI} = \frac{2}{N \cdot (N-1)} \cdot \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} PMI(w_i, w_j)$$

Para cada par de palavras, a *PMI* é calculada, sendo que a média aritmética desses

valores é o resultado dessa métrica. Utilizando o tópico $t_1 = \{$*car*, *driver*, *wheel*, *speed*$\}$ como exemplo, a $C_{UCI}(t_1)$ será:

$$C_{UCI}(t_1) = \frac{2}{4 \cdot (4-1)} \cdot (PMI(car, driver) + PMI(car, wheel) + \\ + PMI(car, speed) + PMI(driver, wheel) + \\ + PMI(driver, speed) + PMI(wheel, speed))$$

3.7 *NPMI Coherence*

A métrica *NPMI Coherence* é uma versão aprimorada da C_{UCI}, que utiliza *NPMI* ao invés de *PMI*, e é definida pela seguinte fórmula:

$$C_{NPMI} = \frac{2}{N \cdot (N-1)} \cdot \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} NPMI(w_i, w_j)$$

Utilizando o tópico $t_1 = \{$*car*, *driver*, *wheel*, *speed*$\}$ como exemplo, a $C_{NPMI}(t_1)$ será:

$$C_{NPMI}(t_1) = \frac{2}{4 \cdot (4-1)} \cdot (NPMI(car, driver) + NPMI(car, wheel) + \\ + NPMI(car, speed) + NPMI(driver, wheel) + \\ + NPMI(driver, speed) + NPMI(wheel, speed))$$

3.8 *UMass Coherence*

A métrica *UMass Coherence* é uma modificação da versão original proposta por MIMNO et al. (2011), e é dada pela seguinte fórmula:

$$C_{Umass} = \frac{2}{N \cdot (N-1)} \cdot \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \left(\frac{P(w_i, w_j) + \epsilon}{P(w_j)} \right)$$

A ideia principal dessa métrica é que a ocorrência de cada palavra dentre as *N-top words* de um tópico deve ser suportada pelas palavras que a precedem (com maior probabilidade). Dessa forma, a probabilidade de uma palavra ocorrer deve ser maior se um documento já contiver uma palavra de ordem mais alta do mesmo tópico. Para cada palavra, o logaritmo de sua probabilidade condicional é calculado a partir de todas as outras palavras que têm uma ordem maior na classificação do tópico como condição. As probabilidades condicionais únicas são

resumidas usando a média aritmética. A constante ϵ pode ser utilizada para evitar a ocorrência de logaritmo de zero.

Utilizando o tópico $t_1 = \{$*car*, *driver*, *wheel*, *speed*$\}$ como exemplo, a $C_{Umass}(t_1)$ será:

$$C_{Umass}(t_1) = \frac{2}{4 \cdot (4-1)} \cdot (\log(P(driver|car)) + \log(P(wheel|car)) + \\ + \log(P(wheel|driver)) + \log(P(speed|car)) + \\ + \log(P(speed|driver)) + \log(P(speed|wheel)))$$

3.9 *C_A Coherence*

Essa métrica mede a associatividade de todos os possíveis pares de palavras de um tópico utilizando uma variação da *NPMI* a partir de uma *context window* de tamanho 5. Essa variação da *NPMI* utiliza o γ para atribuir um peso maior para as maiores associatividades, e ϵ que pode ser usado para evitar o logaritmo de zero:

$$v_{\text{car,driver}} = NPMI(car, driver)^{\gamma} = \left(\frac{PMI(car, driver)}{-\log(P(car, driver) + \epsilon)} \right)^{\gamma}$$

Esse cálculo, após efetuado com todas as palavras, gera um conjunto de vetores, sendo que cada palavra do tópico terá um vetor com o cálculo da *NPMI* para todas as palavras do tópico como elementos:

$$\vec{v}_{\text{car}} = \{NPMI(car, car)^{\gamma}, NPMI(car, driver)^{\gamma}, \\ NPMI(car, wheel)^{\gamma}, NPMI(car, speed)^{\gamma}\}$$

Dessa forma, é calculada a similaridade através do cosseno entre os vetores de cada par de palavras do tópico, e a média aritmética dessas similaridades é o resultado da métrica:

$$C_A = \frac{1}{6} \cdot ((\cos(\vec{v}_{\text{car}}, \vec{v}_{\text{driver}}) + \cos(\vec{v}_{\text{car}}, \vec{v}_{\text{wheel}}) + \cos(\vec{v}_{\text{car}}, \vec{v}_{\text{speed}}) + \\ + \cos(\vec{v}_{\text{driver}}, \vec{v}_{\text{wheel}}) + \cos(\vec{v}_{\text{driver}}, \vec{v}_{\text{speed}}) + \cos(\vec{v}_{\text{wheel}}, \vec{v}_{\text{speed}}))$$

3.10 *C_V Coherence*

Essa métrica mede a coerência de um tópico utilizando uma variação da *NPMI* a partir de uma *sliding window* de tamanho 110, calculando a co-ocorrência de cada palavra do tópico

com todas as palavras desse tópico. Essa variação da *NPMI* utiliza o γ para atribuir um peso maior as associatividades, e ϵ que pode ser usado para evitar o logaritmo de zero:

$$v_{\text{car,driver}} = NPMI(car, driver)^{\gamma} = \left(\frac{PMI(car, driver)}{-\log\left(P(car, driver) + \epsilon\right)} \right)^{\gamma}$$

Esse cálculo, após efetuado com todas as palavras, gera um conjunto de vetores, sendo que cada palavra do tópico terá um vetor com o cálculo da *NPMI* para todas as palavras do tópico como elementos:

$$\vec{v}_{\text{car}} = \{NPMI(car, car)^{\gamma}, NPMI(car, driver)^{\gamma},$$
$$NPMI(car, wheel)^{\gamma}, NPMI(car, speed)^{\gamma}\}$$

Dessa forma, é calculada a similaridade através do cosseno entre o vetor de cada palavra com o vetor resultante da soma dos vetores de cada palavra do tópico. A média aritmética dessas similaridades é o resultado da métrica:

$$\vec{v}_c = \vec{v}_{\text{car}} + \vec{v}_{\text{driver}} + \vec{v}_{\text{wheel}} + \vec{v}_{\text{speed}}$$
$$C_V = \frac{1}{4} \cdot \left(\cos\left(\vec{v}_{\text{car}}, \vec{v}_c\right) + \cos\left(\vec{v}_{\text{driver}}, \vec{v}_c\right) + \cos\left(\vec{v}_{\text{wheel}}, \vec{v}_c\right) + \cos\left(\vec{v}_{\text{speed}}, \vec{v}_c\right)\right)$$

3.11 *C_P Coherence*

Essa métrica mede a coerência de um tópico utilizando a *Confirmation Measure of Fitelson's Coherence* a partir de uma *sliding window* de tamanho 70, e é dada pela seguinte fórmula:

$$C_P = \frac{2}{N \cdot (N-1)} \cdot \sum_{i=2}^{N} \sum_{j=1}^{i-1} m_f(w_i, w_j)$$

Para cada palavra do tópico, é calculada a m_f entre a palavra e as palavras que a precedem na ordem de importância do tópico, o resultado é dado pela média aritmética desses cálculos. O exemplo a seguir mostra a aplicação da métrica para o tópico t_1:

$$C_P(t_1) = \frac{2}{4 \cdot (4-1)} \cdot (m_f(driver, car) + m_f(wheel, car) +$$
$$+ m_f(wheel, driver) + m_f(speed, car) +$$
$$+ m_f(speed, driver) + m_f(speed, wheel))$$

3.12 *Palmetto* - Ferramenta de Avaliação de Qualidade para Tópicos

Palmetto[1] é uma ferramenta de medição da qualidade de tópicos proposta por RÖDER; BOTH; HINNEBURG (2015) que visa ajudar os pesquisadores, oferecendo diferentes cálculos de coerência para as principais palavras de um tópico. Tais métricas de coerência são baseadas nas co-ocorrências de palavras na versão em inglês da *Wikipédia*, e foram comprovadas como correlacionadas através de classificações humanas. A ferramenta possui uma versão *DEMO*[2] que pode ser acessada e utilizada online, e também uma versão de código aberto disponibilizada pelos criadores em um repositório do *Github*[3].

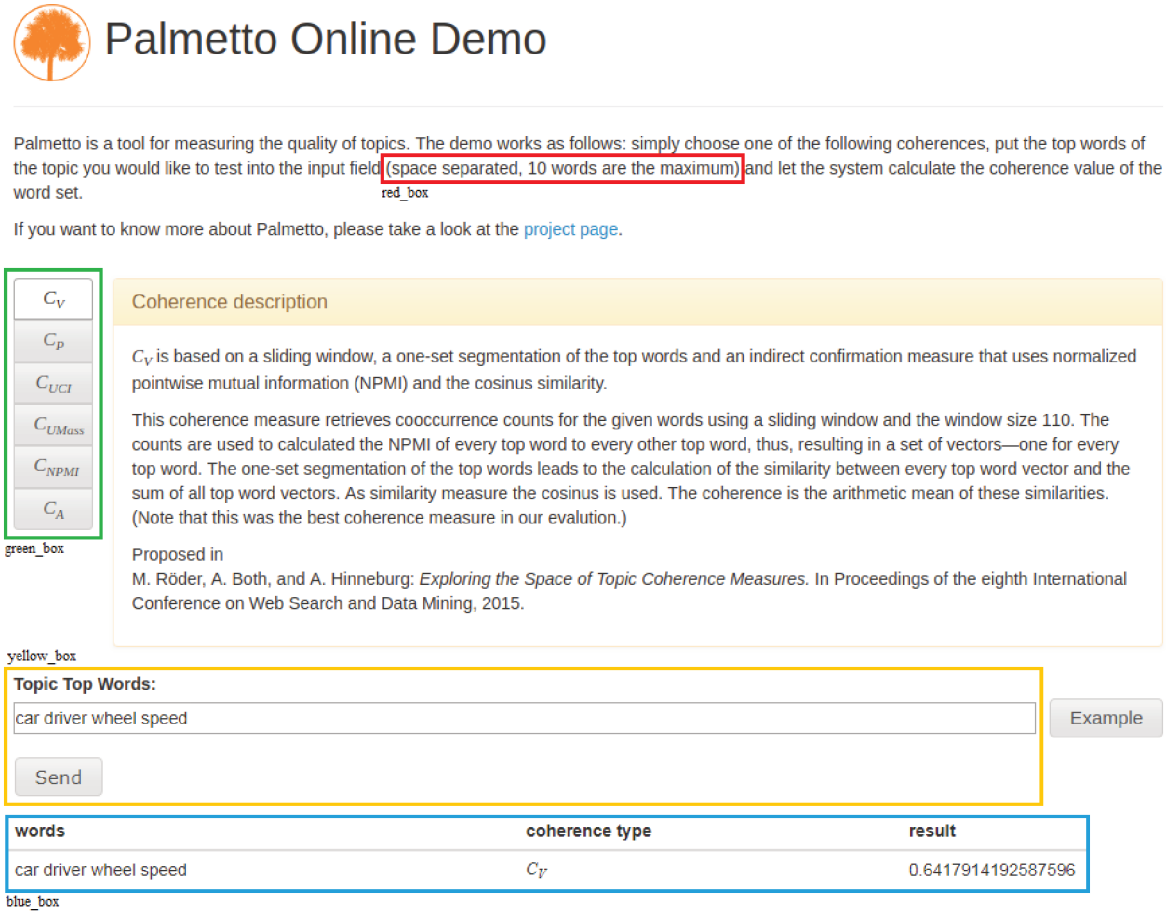

Figura 3.3 – Versão DEMO da ferramenta Palmetto.

A Figura 3.3 mostra um exemplo de utilização da versão DEMO da ferramenta. Destacado em verde (*green_box*) está o menu de seleção da métrica desejada, que nesse caso foi a C_V. Destacado em amarelo (*yellow_box*) está a caixa de texto utilizada para inserção do tópico de-

[1] http://aksw.org/Projects/Palmetto.html
[2] http://palmetto.aksw.org/palmetto-webapp
[3] https://github.com/dice-group/Palmetto

sejado, que tem um limite máximo de 10 palavras, conforme destacado em vermelho (*red_box*). Uma vez selecionada a métrica desejada e inserido o tópico que se quer avaliar, basta clicar no botão *Send* e o resultado da métrica para o tópico será calculado e exibido na mesma tela, conforme destacado em azul (*blue_box*).

4 PROJETO DE EXPERIMENTO

Os experimentos foram realizados conforme a seguinte ordem de execução: definição dos assuntos para a criação de cada tópico; criação dos tópicos por meio da escolha das palavras que compõe cada um deles; validação dos tópicos criados através de formulários; e por fim a aplicação das métricas em cada um dos tópicos. O assunto *Sports* foi usado como teste piloto para a aplicação desta metodologia. O idioma usado para a realização dos experimentos deste trabalho foi o Inglês. Cada um desses passos será explicado de forma completa no decorrer deste capítulo.

4.1 Definição dos Assuntos

Este trabalho adota como estratégia a criação manual dos tópicos, uma vez que o objetivo é avaliar a sensibilidade das métricas, ou seja, nenhum algoritmo de modelagem de tópicos foi utilizado. Para isso, foram definidos cinco assuntos dos quais as palavras que compõe cada tópico foram selecionadas.

Os assuntos foram escolhidos a partir de um único critério: ser de fácil e comum entendimento, ou seja, que são facilmente reconhecidos pela maioria das pessoas, sem a necessidade de pesquisa ou aprofundamento técnico ou teórico. Foram priorizados assuntos que se fazem presentes no dia a dia das pessoas, em jornais, telejornais, revistas e rodas de conversa. A partir deste critério, os cinco assuntos selecionados foram: esportes (*sports*), política (*politics*), religião (*religion*), música (*music*) e natal (*christmas*).

4.2 Criação dos Tópicos

Com os assuntos já definidos, se deu sequência à escolha das palavras que compõe cada um dos tópicos. Para isso, um algoritmo de contagem de palavras foi aplicado, para cada assunto, em uma base de dados contendo dez notícias em inglês sobre o mesmo. Foram escolhidos duas bases de dados (sites web): O site de notícias norte-americano *The New York Times*[4] e a página em inglês sobre o assunto na *Wikipédia*[5]. Em seguida foram escolhidas, por ordem de ocorrência, as palavras que compõe cada um dos tópicos.

[4] https://www.nytimes.com
[5] https://en.wikipedia.org/wiki/Main_Page

4.2.1 Base de Dados

Para cada um dos cinco assuntos, dez entre as notícias mais recentes foram escolhidas através da ferramenta de busca dos sites do *The New York Times*, além da sua página em inglês na *Wikipédia*. Todo esse conteúdo foi armazenado em forma de texto, sendo que a base de dados foi organizada por assunto, e para cada assunto foi organizada em notícias (arquivos de texto de um a dez) e wiki (somente um arquivo). A Figura 4.1 representa a organização da base de dados para o assunto *Sports*.

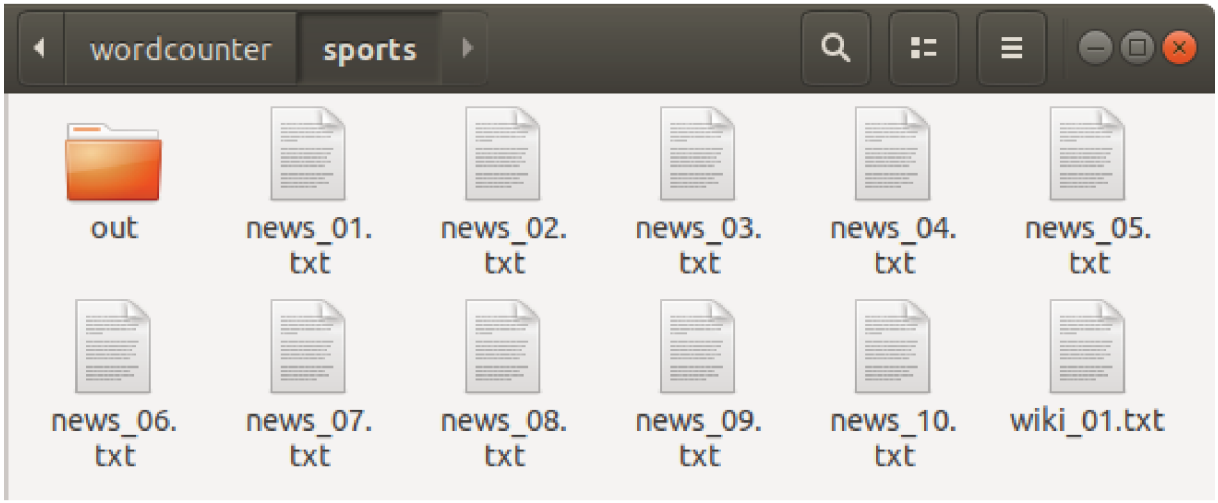

Figura 4.1 – Organização da base de dados do assunto *Sports*.

A escolha do site de notícias norte-americano *The New York Times* se deu pelos seguintes motivos: ser um site de notícias em inglês; ser um dos jornais mais conhecidos e consumidos nos Estados Unidos da América; e por fornecer uma quota de leitura online gratuita diariamente. Já a escolha pela *Wikipédia* se deve ao fato de ser uma enciclopédia online gratuita mundialmente conhecida e por possuir uma versão inteiramente em inglês.

4.2.2 Algoritmo de Contagem de Palavras

O algoritmo de contagem de palavras foi desenvolvido em Python *(versão 3)*, utilizando a biblioteca *NLTK (Natural Lenguage Toolkit)*[6]. Também foram utilizadas algumas bibliotecas auxiliares, como a biblioteca *os* e as bibliotecas *collections* e *operator*. O algoritmo consiste na varredura da base de dados já preparada e organizada previamente, fazendo assim uma contagem de ocorrência de palavras.

[6] https://www.nltk.org/

Após a leitura da entrada (base de dados) e antes de realizar a contagem das palavras, o algoritmo realiza um preprocessamento nos dados. O primeiro passo do algoritmo é realizar a remoção de símbolos e caracteres especiais do corpo do texto. Feito isso, a *tokenização* da entrada é realizada (através do método *word_tokenize*), seguida da remoção das *stop-words* (através da lista de *stop-words* em inglês da biblioteca *NLTK*) e da *stemmização* das palavras (através do método *PorterStemmer*). Por fim, são criados três dicionários contendo a contagem de palavras das notícias, da página da wiki e da junção das notícias com a página da wiki. A saída do algoritmo são três arquivos de texto para cada assunto, contendo os dicionários de palavras ordenados pela ocorrência das mesmas na base de dados. O teto da média das ocorrências também é salvo nestes arquivos. A Figura 4.2 representa a saída do algoritmo para o assunto *Sports*, e a Figura 4.3 representa o arquivo de saída contendo a contagem das palavras, tanto das notícias quanto da página da wiki, para o assunto *Sports*.

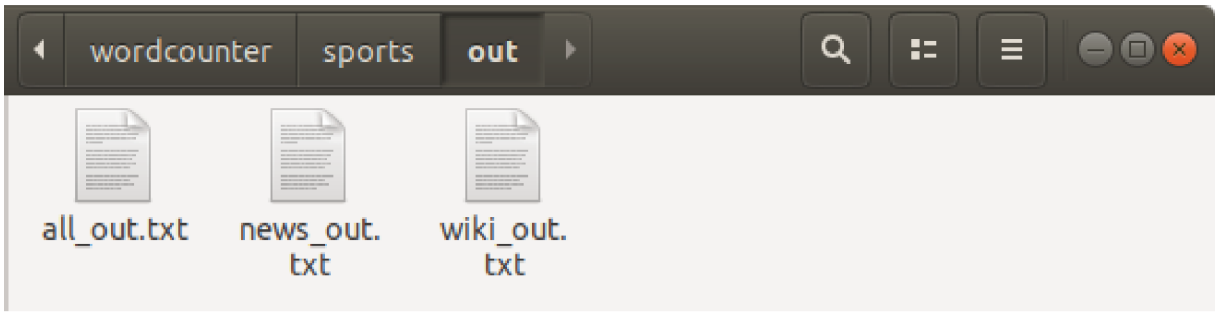

Figura 4.2 – Arquivos de saída do algoritmo de contagem de palavras para o assunto *Sports*.

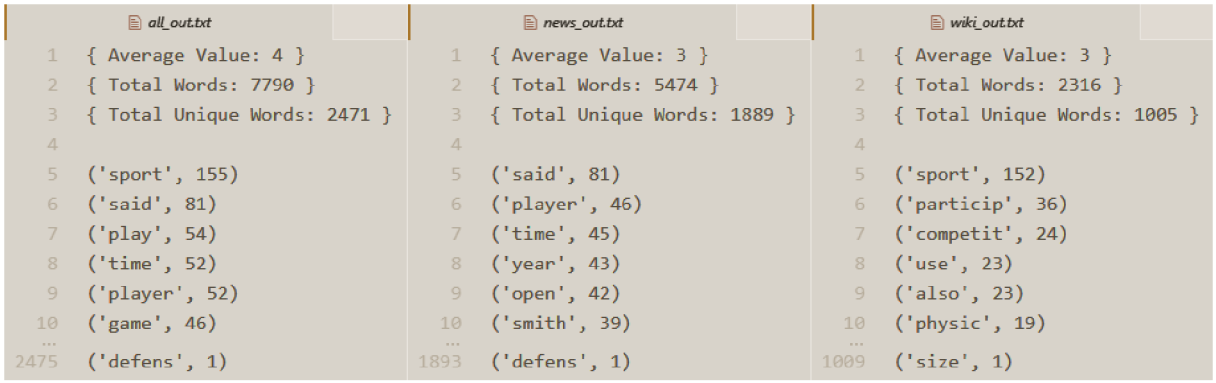

Figura 4.3 – Exemplo de saída do algoritmo de contagem de palavras para o assunto *Sports*.

4.2.3 Escolha das Palavras

A escolha das palavras que compõe cada um dos tópicos foi feita a partir dos arquivos de saída do algoritmo de contagem de palavras, de acordo com os seguintes critérios: número de

ocorrências da palavra e significância da palavra para o determinado assunto, ou seja, palavras com pouca relevância foram desconsideradas. Usando como exemplo o assunto *Sports*, as palavras *player* e *game*, com ocorrências 52 e 46, respectivamente, foram selecionadas. Já as palavras *said* e *open*, com ocorrências 81 e 43, respectivamente, foram desconsideradas. Desta forma, foram escolhidas dez palavras para cada assunto, que foram organizadas em ordem crescente segundo suas ocorrências.

4.2.4 Tópicos Criados

Para cada assunto definido, foram criados um tópico de tamanho 5 (contendo as 5 palavras com maior ocorrência dentre as selecionadas) e um tópico de tamanho 10 (contendo as 10 palavras selecionadas). A Tabela 4.1 apresenta todas as palavras selecionadas para cada assunto, ordenadas por ocorrência.

Sports	**Politics**	**Religion**	**Music**	**Christmas**
player	govern	belief	play	celebration
game	president	church	song	holiday
match	election	moral	instrument	gift
competition	public	god	sound	santa
team	constitution	tradition	composition	december
tournament	corruption	faith	melody	decoration
league	congress	spiritual	concert	tree
athlete	campaign	sacred	singer	birth
score	senator	holy	symphony	jesus
goal	law	supernatural	rhythm	feast

Tabela 4.1 – Palavras selecionadas para compor os tópicos de cada assunto.

Tomando o assunto *Sports* como exemplo, os dois tópicos criados foram: $Sports_{top5} = \{player, game, match, competition, team\}$ e $Sports_{top10} = \{player, game, match, competition, team, tournament, league, athlete, score, goal\}$.
A linha pontilhada na tabela delimita os tamanhos dos tópicos: as cinco palavras acima da linha pontilhada constituem os tópicos de tamanho 5, e as dez palavras, tanto acima quanto abaixo da linha pontilhada, constituem os tópicos de tamanho 10. Cada assunto gerou dois tópicos com esses formatos.

4.3 Método de Palavras Intrusas

Para criar variações dos tópicos gerados a partir dos assuntos selecionados, a técnica de inserção de palavras intrusas (CHANG et al., 2009) foi utilizada. Essa técnica consiste na troca de palavras de um tópico por palavras que não estejam associadas ao mesmo, de forma que essas palavras sejam facilmente identificadas como intrusas. Como exemplo, o tópico $Animais_{top5} = \{cachorro, gato, cavalo, porco, vaca\}$ teve a palavra *cavalo* trocada pela palavra *carro*, gerando o tópico $Animais_{top5} = \{cachorro, gato, carro, porco, vaca\}$. Colocando todas as palavras do novo tópico em um contexto, é fácil identificar que a palavra *carro* não pertence ao mesmo.

Foram criadas duas variações para cada tópico de tamanho 5, e quatro variações para cada tópico de tamanho 10. As variações dos tópicos de tamanho 5 foram criadas da seguinte forma: uma palavra com ocorrência dentro da média (*palavraM1*) e uma palavra com ocorrência igual a um (*palavraU1*) foram selecionadas e trocadas pelas últimas palavras do tópico original, conforme representado na Tabela 4.2.

Original	**1 Intrusa**	**2 Intrusas**
palavra1	palavra1	palavra1
palavra2	palavra2	palavra2
palavra3	palavra3	palavra3
palavra4	palavra4	palavraM1
palavra5	palavraM1	palavraU1

Tabela 4.2 – Inserção de palavras intrusas nos tópicos de tamanho 5.

As variações dos tópicos de tamanho 10 foram criadas da seguinte forma: duas palavras com ocorrência dentro da média (*palavraM1* e *palavraM2*) e duas palavras com ocorrência igual a um (*palavraU1* e *palavraU2*) foram selecionadas e trocadas pelas últimas palavras do tópico original, conforme representado na Tabela 4.3.

A escolha das palavras intrusas como sendo palavras de ocorrência dentro da média e palavras de ocorrência igual a um se deu a partir da seguinte ideia: as palavras de ocorrência dentro da média são palavras que não remetem diretamente ao assunto, mas que podem ser associadas ao mesmo; já as palavras com ocorrência igual a um são palavras que não remetem diretamente ao assunto, além de serem difíceis de serem associadas ao mesmo. Desta forma, os tópicos foram deturpados com palavras intrusas fracas e palavras intrusas fortes.

Sendo assim, cada assunto gerou um total de oito tópicos, sendo três tópicos de tamanho

Original	1 Intrusa	2 Intrusas	3 Intrusas	4 Intrusas
palavra1	palavra1	palavra1	palavra1	palavra1
palavra2	palavra2	palavra2	palavra2	palavra2
palavra3	palavra3	palavra3	palavra3	palavra3
palavra4	palavra4	palavra4	palavra4	palavra4
palavra5	palavra5	palavra5	palavra5	palavra5
palavra6	palavra6	palavra6	palavra6	palavra6
palavra7	palavra7	palavra7	palavra7	palavraM1
palavra8	palavra8	palavra8	palavraM1	palavraU1
palavra9	palavra9	palavraM1	palavraU1	palavraM2
palavra10	palavraM1	palavraU1	palavraM2	palavraU2

Tabela 4.3 – Inserção de palavras intrusas nos tópicos de tamanho 10.

5 e cinco tópicos de tamanho 10. Todos os tópicos serão apresentados na próxima seção, e serão nomeados da seguinte forma: $Assunto_{i-j}$, onde i é o tamanho do tópico e j é a quantidade de palavras intrusas no mesmo.

4.4 Tópicos Gerados

Nesta seção serão apresentados todos os 40 tópicos gerados pelos procedimentos das seções anteriores.

Os tópicos sobre esportes (*sports*) estão apresentados nas Tabelas 4.4 e 4.5, e tem como intrusas as palavras: *watch* e *push* como palavras na média de ocorrência, e *tree* e *mirror* como palavras com ocorrência igual a um.

$Sports_{5-0}$	$Sports_{5-1}$	$Sports_{5-2}$
player	player	player
game	game	game
match	match	match
competition	competition	watch
team	watch	tree

Tabela 4.4 – Tópicos de tamanho 5 sobre o assunto Esportes (*Sports*).

Os tópicos sobre política (*politics*) estão apresentados nas Tabelas 4.6 e 4.7, e tem como intrusas as palavras: *media* e *money* como palavras na média de ocorrência, e *steel* e *cloud* como palavras com ocorrência igual a um.

Os tópicos sobre religião (*religion*) estão apresentados nas Tabelas 4.8 e 4.9, e tem como intrusas as palavras: *trade* e *paper* como palavras na média de ocorrência, e *island* e *fuel* como palavras com ocorrência igual a um.

$Sports_{10-0}$	$Sports_{10-1}$	$Sports_{10-2}$	$Sports_{10-3}$	$Sports_{10-4}$
player	player	player	player	player
game	game	game	game	game
match	match	match	match	match
competition	competition	competition	competition	competition
team	team	team	team	team
tournament	tournament	tournament	tournament	tournament
league	league	league	league	watch
athlete	athlete	athlete	watch	tree
score	score	watch	tree	push
goal	watch	tree	push	mirror

Tabela 4.5 – Tópicos de tamanho 10 sobre o assunto Esportes (*Sports*).

$Politics_{5-0}$	$Politics_{5-1}$	$Politics_{5-2}$
govern	govern	govern
president	president	president
election	election	election
public	public	media
constitution	media	steel

Tabela 4.6 – Tópicos de tamanho 5 sobre o assunto Política (*Politics*).

Os tópicos sobre música (*music*) estão apresentados nas Tabelas 4.10 e 4.11, e tem como intrusas as palavras: *contrast* e *purpose* como palavras na média de ocorrência, e *farm* e *plane* como palavras com ocorrência igual a um.

Por último, os tópicos sobre natal (*christmas*) estão apresentados nas Tabelas 4.12 e 4.13, e tem como intrusas as palavras: *case* e *plastic* como palavras na média de ocorrência, e *grass* e *sea* como palavras com ocorrência igual a um.

4.5 Validação dos Tópicos

Com o intuito de verificar se os tópicos criados representam de fato os assuntos que os geraram, uma técnica de validação a partir de formulários foi utilizada. Os formulários criados tiveram como objetivo atestar que pessoas conseguem identificar o assunto apenas observando as palavras que compõe os tópicos, e, se obtendo sucesso, qualificar os tópicos criados como ótimos. Para isso, foram criados dois tipos de formulários: um contendo os tópicos originais (sem nenhuma palavra intrusa), no qual o objetivo é classificar o tópico atribuindo um assunto ao mesmo, e outro contendo os tópicos com suas palavras intrusas, onde o objetivo é identificar tais palavras. Os formulários foram elaborados através da plataforma de criação de formulários

$Politics_{10-0}$	$Politics_{10-1}$	$Politics_{10-2}$	$Politics_{10-3}$	$Politics_{10-4}$
govern	govern	govern	govern	govern
president	president	president	president	president
election	election	election	election	election
public	public	public	public	public
constitution	constitution	constitution	constitution	constitution
corruption	corruption	corruption	corruption	corruption
congress	congress	congress	congress	media
campaign	campaign	campaign	media	steel
senator	senator	media	steel	money
law	media	steel	money	cloud

Tabela 4.7 – Tópicos de tamanho 10 sobre o assunto Política (*Politics*).

$Religion_{5-0}$	$Religion_{5-1}$	$Religion_{5-2}$
belief	belief	belief
church	church	church
moral	moral	moral
god	god	trade
tradition	trade	island

Tabela 4.8 – Tópicos de tamanho 5 sobre o assunto Religião (*Religion*).

da *Google*[7]. Todas as respostas foram obtidas de forma anônima, ou seja, os pesquisados não foram identificados ou associados a nenhuma das respostas.

4.5.1 Formulário de Classificação

O formulário de classificação de tópicos é composto por cinco perguntas (uma para cada assunto), sendo que cada pergunta contém uma lista com as dez palavras do tópico (nenhuma palavra intrusa) do referido assunto. Cada uma dessas perguntas foi respondida com um texto curto, associando um assunto que melhor represente as palavras listadas. Por fim, uma sexta pergunta de múltipla escolha para avaliar a dificuldade encontrada em responder as questões anteriores, com valor de 1 a 5, sendo: 1 para muito fácil, 2 para fácil, 3 para regular, 4 para difícil e 5 para muito difícil. A Figura 4.4 representa o formulário de classificação, utilizando o assunto *Sports* como exemplo (somente a primeira e a sexta pergunta estão representadas na figura).

[7] https://www.google.com/forms/about/

$Religion_{10-0}$	$Religion_{10-1}$	$Religion_{10-2}$	$Religion_{10-3}$	$Religion_{10-4}$
belief	belief	belief	belief	belief
church	church	church	church	church
moral	moral	moral	moral	moral
god	god	god	god	god
tradition	tradition	tradition	tradition	tradition
faith	faith	faith	faith	faith
spiritual	spiritual	spiritual	spiritual	trade
sacred	sacred	sacred	trade	island
holy	holy	trade	island	paper
supernatural	trade	island	paper	fuel

Tabela 4.9 – Tópicos de tamanho 10 sobre o assunto Religião (*Religion*).

$Music_{5-0}$	$Music_{5-1}$	$Music_{5-2}$
play	play	play
song	song	song
instrument	instrument	instrument
sound	sound	contrast
composition	contrast	farm

Tabela 4.10 – Tópicos de tamanho 5 sobre o assunto Música (*Music*).

4.5.2 Formulários de Palavras Intrusas

Foi criado um formulário para cada configuração de tópico, totalizando 6 formulários: um formulário para todos os tópicos de tamanho 5 com uma palavra intrusa; um formulário para todos os tópicos de tamanho 5 com duas palavras intrusas; um formulário para todos os tópicos de tamanho 10 com uma palavra intrusa; um formulário para todos os tópicos de tamanho 10 com duas palavras intrusas; um formulário para todos os tópicos de tamanho 10 com três palavras intrusas; e um formulário para todos os tópicos de tamanho 10 com quatro palavras intrusas.

Cada formulário é composto por cinco perguntas (uma para cada assunto) apresentando a lista de palavras que compõe o determinado tópico e o número de palavras intrusas no mesmo, onde o objetivo é identificar quais palavras não pertencem àquele tópico, ou seja, identificar as palavras intrusas. Por fim, uma sexta pergunta de múltipla escolha para avaliar a dificuldade encontrada em responder as questões anteriores, com valor de 1 a 5, sendo: 1 para muito fácil, 2 para fácil, 3 para regular, 4 para difícil e 5 para muito difícil. A Figura 4.5 representa o formulário de palavras intrusas para tópicos de tamanho 5 com duas palavras intrusas, e utiliza o assunto *Sports* como exemplo (somente a primeira e a sexta pergunta estão representadas na

$Music_{10-0}$	$Music_{10-1}$	$Music_{10-2}$	$Music_{10-3}$	$Music_{10-4}$
play	play	play	play	play
song	song	song	song	song
instrument	instrument	instrument	instrument	instrument
sound	sound	sound	sound	sound
composition	composition	composition	composition	composition
melody	melody	melody	melody	melody
concert	concert	concert	concert	contrast
singer	singer	singer	contrast	farm
symphony	symphony	contrast	farm	purpose
rhythm	contrast	farm	purpose	plane

Tabela 4.11 – Tópicos de tamanho 10 sobre o assunto Música (*Music*).

$Christmas_{5-0}$	$Christmas_{5-1}$	$Christmas_{5-2}$
celebration	celebration	celebration
holiday	holiday	holiday
gift	gift	gift
santa	santa	case
december	case	grass

Tabela 4.12 – Tópicos de tamanho 5 sobre o assunto Natal (*Christmas*).

figura). As palavras foram embaralhadas e tiradas de ordem, evitando assim alguma sugestão devido à posição das palavras na lista.

4.5.3 Perfil das Pessoas Pesquisadas

Para realizar a validação dos tópicos de forma controlada, todas as pessoas pesquisadas se enquadram no seguinte perfil: possuir ou estar cursando o ensino superior e ter um bom nível de conhecimento da língua inglesa. No total, foram selecionadas 63 pessoas para responder a pesquisa, sendo que estas foram divididas em 7 grupos distintos. Cada grupo recebeu um dos 7 formulários para responder, sendo que em nenhum momento tiveram acesso aos demais, evitando que as respostas fossem influenciadas pelas informações contidas nos outros formulários.

4.5.4 Resultados dos Formulários

Nesta subseção serão apresentados os resultados obtidos pela pesquisa de avaliação dos tópicos através dos formulários de classificação e palavras intrusas, sendo que o assunto *Sports* foi utilizado como teste piloto (processo explicado na seção 4.6), ou seja, foi aplicado de forma individual antes dos demais assuntos, e por isso possui uma avaliação individual de dificuldade.

A Tabela 4.14 apresenta os resultados obtidos pela aplicação do formulário de classifi-

$Christmas_{10-0}$	$Christmas_{10-1}$	$Christmas_{10-2}$	$Christmas_{10-3}$	$Christmas_{10-4}$
celebration	celebration	celebration	celebration	celebration
holiday	holiday	holiday	holiday	holiday
gift	gift	gift	gift	gift
santa	santa	santa	santa	santa
december	december	december	december	december
decoration	decoration	decoration	decoration	decoration
tree	tree	tree	tree	case
birth	birth	birth	case	grass
jesus	jesus	case	grass	plastic
feast	case	grass	plastic	sea

Tabela 4.13 – Tópicos de tamanho 10 sobre o assunto Natal (*Christmas*).

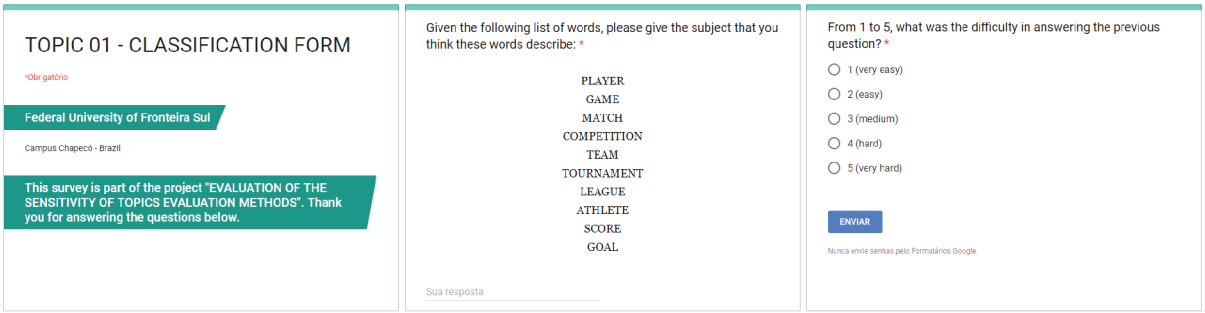

Figura 4.4 – Exemplo de formulário de classificação.

cação. É possível constatar que todas as respostas estão diretamente relacionadas ao seu devido assunto. É possível constatar também que a dificuldade encontrada pelos pesquisados para responder as questões varia entre muito fácil e difícil, sendo que a maioria declarou entre fácil e muito fácil.

RESULTADOS DO FORMULÁRIO DE CLASSIFICAÇÃO

SPORTS		POLITICS		RELIGION		MUSIC		CHRISTMAS	
Resposta	QTD	Resposta	QTD	Resposta	QTD	Resposta	QTD	Resposta	QTD
Games	3	Politics	4	Religion	5	Music	4	Christmas	5
Sports	4	Communism	1			Band	1		
Soccer	1								

DIFICULDADE								DIFICULDADE	
Very Easy	3							Very Easy	2
Easy	3							Easy	2
Regular	1							Regular	0
Hard	1							Hard	1
Very Hard	0							Very Hard	0

Tabela 4.14 – Resultado do formulário de classificação.

As Tabelas 4.15 e 4.16 apresentam os resultados obtidos pela pesquisa de avaliação de tópicos através dos formulários de palavras intrusas, para tópicos de tamanho 5, sendo que o assunto *Sports* foi utilizado como teste piloto. É possível constatar que para estas duas configurações de tópicos os resultados obtidos foram considerados excelentes. De modo geral, as

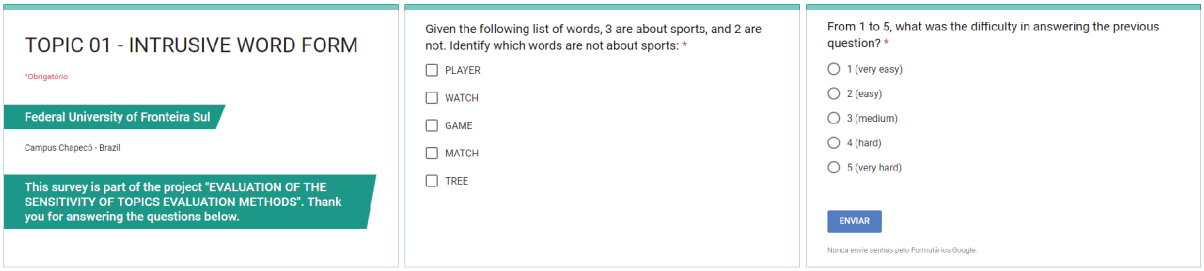

Figura 4.5 – Exemplo de formulário de palavras intrusas.

palavras intrusas foram identificadas e a maioria dos pesquisados declararam a dificuldade em responder como sendo entre muito fácil e regular.

RESULTADOS DO FORMULÁRIO DE PALAVRAS INTRUSAS PARA TÓPICOS DE TAMANHO 5 COM UMA PALAVRA INTRUSA

SPORTS		POLITICS		RELIGION		MUSIC		CHRISTMAS	
Palavra	QTD	Palavra	QTD	Palavra	QTD	Palavra	QTD	Palavra	QTD
player	0	govern	0	belief	0	play	0	celebration	0
game	1	president	0	church	0	song	0	holiday	0
match	0	election	0	moral	0	instrument	0	gift	0
competition	1	public	1	god	0	sound	0	santa	0
watch	11	media	6	trade	7	contrast	7	case	7
DIFICULDADE								**DIFICULDADE**	
Very Easy	0							Very Easy	1
Easy	6							Easy	3
Regular	5							Regular	3
Hard	1							Hard	0
Very Hard	1							Very Hard	0

Tabela 4.15 – Resultado do formulário de palavras intrusas 4/1.

RESULTADOS DO FORMULÁRIO DE PALAVRAS INTRUSAS PARA TÓPICOS DE TAMANHO 5 COM DUAS PALAVRAS INTRUSAS

SPORTS		POLITICS		RELIGION		MUSIC		CHRISTMAS	
Palavra	QTD	Palavra	QTD	Palavra	QTD	Palavra	QTD	Palavra	QTD
player	0	govern	0	belief	0	play	0	celebration	0
game	0	president	0	church	0	song	0	holiday	0
match	1	election	0	moral	0	instrument	0	gift	0
watch	7	media	6	trade	6	contrast	6	case	6
tree	8	steel	6	island	6	farm	6	grass	6
DIFICULDADE								**DIFICULDADE**	
Very Easy	5							Very Easy	4
Easy	3							Easy	1
Regular	0							Regular	1
Hard	0							Hard	0
Very Hard	0							Very Hard	0

Tabela 4.16 – Resultado do formulário de palavras intrusas 3/2.

As Tabelas 4.17, 4.18, 4.19 e 4.20 apresentam os resultados obtidos pela pesquisa de avaliação de tópicos através dos formulários de palavras intrusas, para tópicos de tamanho 10, sendo que o assunto *Sports* foi utilizado como teste piloto. É possível constatar que para estas quatro configurações de tópicos os resultados obtidos foram considerados bons. De modo geral, as palavras intrusas foram identificadas, embora alguns dos pesquisados tenham identificado algumas palavras de forma errada. Os pesquisados declararam a dificuldade em responder como

sendo entre muito fácil e difícil, embora a grande maioria tenha declarado a dificuldade como entre muito fácil e regular.

RESULTADOS DO FORMULÁRIO DE PALAVRAS INTRUSAS PARA TÓPICOS DE TAMANHO 10 COM UMA PALAVRA INTRUSA

SPORTS		POLITICS		RELIGION		MUSIC		CHRISTMAS	
Palavra	QTD	Palavra	QTD	Palavra	QTD	Palavra	QTD	Palavra	QTD
player	0	govern	0	belief	0	play	0	celebration	0
game	0	president	0	church	0	song	0	holiday	0
match	1	election	0	moral	0	instrument	0	gift	0
competition	0	public	1	god	0	sound	0	santa	0
team	0	constitution	0	tradition	0	composition	0	december	0
tournament	0	corruption	0	faith	0	melody	0	decoration	0
league	0	congress	0	spiritual	0	concert	0	tree	0
athlete	0	campaign	0	sacred	0	singer	0	birth	0
score	0	senator	0	holy	0	symphony	0	jesus	0
watch	7	media	4	trade	5	contrast	5	case	5

DIFICULDADE	
Very Easy	1
Easy	2
Regular	3
Hard	1
Very Hard	1

DIFICULDADE	
Very Easy	0
Easy	2
Regular	3
Hard	0
Very Hard	0

Tabela 4.17 – Resultado do formulário de palavras intrusas 9/1.

RESULTADOS DO FORMULÁRIO DE PALAVRAS INTRUSAS PARA TÓPICOS DE TAMANHO 10 COM DUAS PALAVRAS INTRUSAS

SPORTS		POLITICS		RELIGION		MUSIC		CHRISTMAS	
Palavra	QTD	Palavra	QTD	Palavra	QTD	Palavra	QTD	Palavra	QTD
player	0	govern	0	belief	0	play	0	celebration	0
game	0	president	0	church	0	song	0	holiday	0
match	1	election	0	moral	0	instrument	0	gift	0
competition	0	public	0	god	0	sound	0	santa	0
team	0	constitution	0	tradition	0	composition	0	december	0
tournament	0	corruption	0	faith	0	melody	0	decoration	0
league	0	congress	0	spiritual	0	concert	1	tree	0
athlete	0	campaign	0	sacred	0	singer	0	birth	0
watch	7	media	7	trade	7	contrast	6	case	7
tree	8	steel	7	island	7	farm	7	grass	7

DIFICULDADE	
Very Easy	3
Easy	1
Regular	2
Hard	0
Very Hard	2

DIFICULDADE	
Very Easy	1
Easy	4
Regular	2
Hard	0
Very Hard	0

Tabela 4.18 – Resultado do formulário de palavras intrusas 8/2.

Pode-se constatar que houveram mais respostas para os formulários referentes ao assunto *Sports*, que foi utilizado como teste piloto, porém o número de respostas obtidas para os demais assuntos também foi julgado suficiente. De modo geral, os resultados obtidos pela avaliação de tópicos através de formulários foram considerados bons. Dessa forma, todos os tópicos propostos a partir dos assuntos previamente selecionados foram julgados aptos à serem utilizados nas demais etapas deste trabalho.

RESULTADOS DO FORMULÁRIO DE PALAVRAS INTRUSAS PARA TÓPICOS DE TAMANHO 10 COM TRÊS PALAVRAS INTRUSAS								
SPORTS		POLITICS		RELIGION		MUSIC		CHRISTMAS	
Palavra	QTD	Palavra	QTD	Palavra	QTD	Palavra	QTD	Palavra	QTD
player	0	govern	0	belief	0	play	1	celebration	0
game	0	president	0	church	0	song	0	holiday	0
match	0	election	0	moral	0	instrument	0	gift	0
competition	0	public	0	god	0	sound	0	santa	0
team	0	constitution	0	tradition	0	composition	0	december	0
tournament	0	corruption	0	faith	0	melody	0	decoration	0
league	0	congress	0	spiritual	0	concert	0	tree	0
watch	7	media	6	trade	6	contrast	5	case	6
tree	7	steel	6	island	6	farm	6	grass	6
push	7	money	6	paper	6	purpose	6	plastic	6
DIFICULDADE								DIFICULDADE	
Very Easy	0							Very Easy	0
Easy	6							Easy	5
Regular	0							Regular	0
Hard	1							Hard	1
Very Hard	0							Very Hard	0

Tabela 4.19 – Resultado do formulário de palavras intrusas 7/3.

RESULTADOS DO FORMULÁRIO DE PALAVRAS INTRUSAS PARA TÓPICOS DE TAMANHO 10 COM QUATRO PALAVRAS INTRUSAS								
SPORTS		POLITICS		RELIGION		MUSIC		CHRISTMAS	
Palavra	QTD	Palavra	QTD	Palavra	QTD	Palavra	QTD	Palavra	QTD
player	0	govern	0	belief	0	play	0	celebration	0
game	0	president	0	church	0	song	0	holiday	0
match	1	election	1	moral	0	instrument	0	gift	0
competition	0	public	3	god	0	sound	0	santa	0
team	0	constitution	0	tradition	1	composition	1	december	0
tournament	0	corruption	0	faith	0	melody	0	decoration	0
watch	7	media	7	trade	8	contrast	7	case	8
tree	8	steel	8	island	8	farm	8	grass	8
push	8	money	5	paper	8	purpose	8	plastic	8
mirror	8	cloud	8	fuel	7	plane	8	sea	8
DIFICULDADE								DIFICULDADE	
Very Easy	0							Very Easy	0
Easy	5							Easy	4
Regular	3							Regular	4
Hard	0							Hard	0
Very Hard	0							Very Hard	0

Tabela 4.20 – Resultado do formulário de palavras intrusas 6/4.

4.6 Teste Piloto

Para fins de verificação da viabilidade, todos os processos anteriormente descritos foram realizados uma vez utilizando o assunto *Sports* como teste piloto, desde a definição do assunto até a aplicação dos formulários de avaliação (com os mesmos 7 grupos de pesquisados) e análise dos seus resultados. Após a constatação da viabilidade desses processos, os mesmos foram aplicados aos demais assuntos.

4.7 Aplicação das Métricas

Uma vez que todos os tópicos foram criados e validados, a plataforma online (versão DEMO) do *Palmetto* foi utilizada para a aplicação das métricas. Para isso, foi desenvolvido um

algoritmo em Python *versão 3*, utilizando a biblioteca de automação de browser *selenium*, além do auxílio das bibliotecas de sistema *random*, *time*, *datetime* e *csv*. O algoritmo recebe como entrada um arquivo de texto contendo todos os tópicos, separados por nova linha, e acessa a plataforma do *Palmetto* aplicando todas as métricas a cada um desses tópicos. Por fim, o algoritmo salva os resultados obtidos de cada uma das métricas em um arquivo com extensão *csv* que pode ser aberto em formato de planilha. O algoritmo também mantém um arquivo de *log* contendo informações como: andamento do algoritmo, resultados obtidos e tempo de execução. Os resultados obtidos serão apresentados no Capítulo 5.

5 ANÁLISE DOS RESULTADOS

Este capítulo apresenta os resultados obtidos através da execução dos procedimentos de aplicação das métricas (através da versão DEMO do *Palmetto*) que foram previamente descritos, bem como uma análise gráfica avaliativa em relação a sensibilidade das métricas de acordo com os valores obtidos.

5.1 Resultados da Aplicação das Métricas

Esta seção apresenta, através da Tabela 5.1, os valores obtidos por meio da aplicação das seis métricas propostas por RÖDER; BOTH; HINNEBURG (2015) em cada um dos tópicos criados.

A primeira análise feita diz respeito ao comportamento, de modo geral, de todas as métricas. Os tópicos foram criados e modificados, através da inserção de palavras intrusas, para que as palavras que os compõem tivessem suas relações enfraquecidas. Esse comportamento pode ser atestado mediante a pontuação aferida pelas métricas a cada um dos tópicos. De modo geral, é possível verificar que para todas as métricas, quanto mais palavras intrusas, menor a pontuação do tópico, reforçando as expectativas previamente criadas. Algumas anomalias neste comportamento esperado podem ser observadas em alguns casos como:

- a métrica C_V para os tópicos do assunto *Christmas*;
- a métrica C_{UMASS} para os tópicos de tamanho 10 do assunto *Religion*;
- a métrica C_{UMASS} para os tópicos de tamanho 10 do assunto *Christmas*;
- a métrica C_A para os tópicos de tamanho 5 do assunto *Religion*;
- e a métrica C_A para os tópicos do assunto *Christmas*.

A segunda análise diz respeito aos tópicos que geraram as anomalias descritas na primeira análise. É possível afirmar que tais anomalias na pontuação dos tópicos em questão se devem ao fato de que, de algum modo, mesmo as palavras intrusas sendo identificadas pelos pesquisados, algumas delas ainda possuem algum relacionamento significante quando suas similaridades de co-ocorrência são medidas, ou seja, são palavras que acabam ocorrendo juntas.

Tópico	C_V	C_P	C_{UCI}	C_{UMASS}	C_{NPMI}	C_A
$Sports_{5-0}$	0,61026	0,64951	1,33859	-1,31801	0,15826	0,29836
$Sports_{5-1}$	0,52885	0,41231	0,75038	-2,10428	0,08217	0,17603
$Sports_{5-2}$	0,51842	0,14417	0,17047	-2,50532	0,03231	0,14407
$Sports_{10-0}$	0,56825	0,67819	1,47196	-1,75147	0,16872	0,34443
$Sports_{10-1}$	0,51917	0,55148	1,09217	-1,97090	0,12477	0,27435
$Sports_{10-2}$	0,48026	0,33620	0,47239	-2,32378	0,06856	0,22436
$Sports_{10-3}$	0,46620	0,26386	0,37691	-2,33003	0,05904	0,20449
$Sports_{10-4}$	0,42114	0,16051	0,14701	-2,69445	0,03155	0,15882
$Politics_{5-0}$	0,51630	0,54484	1,06511	-1,70655	0,09762	0,26678
$Politics_{5-1}$	0,48160	0,31313	0,41141	-1,73102	0,04121	0,20960
$Politics_{5-2}$	0,47723	0,03351	-0,30131	-2,60790	-0,01072	0,19147
$Politics_{10-0}$	0,42096	0,53857	1,12538	-1,90088	0,10358	0,26896
$Politics_{10-1}$	0,40425	0,44843	0,87339	-2,14379	0,08046	0,23233
$Politics_{10-2}$	0,38308	0,27215	0,14874	-2,35650	0,03552	0,20646
$Politics_{10-3}$	0,37251	0,22530	-0,02080	-2,39128	0,01972	0,17883
$Politics_{10-4}$	0,35710	0,09095	-0,84230	-2,76313	-0,02279	0,17519
$Religion_{5-0}$	0,51735	0,62138	1,63980	-1,83562	0,14462	0,27495
$Religion_{5-1}$	0,49881	0,30043	0,65993	-1,98137	0,06474	0,21573
$Religion_{5-2}$	0,46766	0,06957	-0,14405	-2,07489	-0,00587	0,23707
$Religion_{10-0}$	0,45566	0,68889	1,96726	-2,32925	0,16619	0,29444
$Religion_{10-1}$	0,44305	0,51472	1,39211	-2,04059	0,12447	0,27436
$Religion_{10-2}$	0,41414	0,36744	0,93599	-2,06367	0,08523	0,24854
$Religion_{10-3}$	0,39629	0,24783	0,54835	-2,02664	0,05400	0,21102
$Religion_{10-4}$	0,37329	0,09187	0,05855	-2,27747	0,01513	0,18709
$Music_{5-0}$	0,51593	0,54569	1,16464	-2,32002	0,11393	0,19962
$Music_{5-1}$	0,49365	0,37182	0,79871	-2,50728	0,07997	0,16224
$Music_{5-2}$	0,46655	0,03799	-0,10555	-2,95242	0,00076	0,09209
$Music_{10-0}$	0,45823	0,63700	1,40555	-2,40416	0,12475	0,21424
$Music_{10-1}$	0,43047	0,51670	1,08273	-2,47470	0,09741	0,18049
$Music_{10-2}$	0,40732	0,32599	0,60823	-2,51469	0,06061	0,16110
$Music_{10-3}$	0,38784	0,23449	0,46663	-2,60157	0,04678	0,14789
$Music_{10-4}$	0,37245	0,14046	0,00293	-2,84725	0,02026	0,13397
$Christmas_{5-0}$	0,47125	0,36788	1,00773	-2,03109	0,08043	0,16102
$Christmas_{5-1}$	0,47195	0,22314	0,60711	-2,25784	0,04942	0,15891
$Christmas_{5-2}$	0,46736	0,09208	-1,01397	-2,73181	-0,02271	0,17143
$Christmas_{10-0}$	0,35272	0,33511	0,83320	-2,82375	0,06492	0,16187
$Christmas_{10-1}$	0,33463	0,18245	0,36802	-2,48817	0,02909	0,13307
$Christmas_{10-2}$	0,33568	0,13705	-0,02681	-2,64255	0,01120	0,13843
$Christmas_{10-3}$	0,33591	0,12885	-0,30633	-2,89929	0,00076	0,14359
$Christmas_{10-4}$	0,33066	0,08298	-0,49598	-2,82356	-0,01389	0,13295

Tabela 5.1 – Resultado da aplicação das métricas através do *Palmetto*.

5.2 Análises dos Valores

Esta seção apresenta, através de gráficos e tabelas, o comportamento de cada uma das métricas. Todos os gráficos apresentados foram gerados a partir de um algoritmo que recebe as pontuações dos tópicos para cada métrica como entrada, e que foi desenvolvido em python (*versão 3*) com auxílio da biblioteca gráfica *MatPlotLib*[8]. Todas as tabelas de sensibilidade foram geradas através do cálculo da sensibilidade descrito na subseção 5.2.2.

5.2.1 Gráficos Gerados

Os gráficos gerados a partir dos valores obtidos pela aplicação das métricas foram separados pelo assunto dos tópicos. Para cada assunto, dois gráficos são apresentados, sendo o gráfico da esquerda referente aos tópicos de tamanho 5 e o gráfico da direita referente aos tópicos de tamanho 10. O eixo X dos gráficos representa o número de palavras intrusas no tópico, e o eixo Y representa a pontuação obtida pelo tópico na referida métrica. Cada métrica está apresentada de uma cor diferente.

Todos os valores da métrica C_{UMASS} foram alterados para o respectivo valor absoluto, a fim de melhor aproveitar o enquadramento do gráfico. Devido a isto, a interpretação dos valores para esta métrica nos gráficos deve ser a seguinte: quanto maior o valor, pior a pontuação do tópico. Os demais valores das demais métricas foram mantidos inalterados.

5.2.2 Sensibilidade das Métricas

As Tabelas 5.2, 5.3, 5.4, 5.5 e 5.6 apresentam a sensibilidade de cada uma das métricas para os tópicos do determinado assunto. As tabelas estão divididas por uma linha tracejada que divide a sensibilidade calculada para os tópicos de tamanho 5 e para os tópicos de tamanho 10. Para cada tamanho de tópico são apresentadas as médias de variação absoluta e a porcentagem em relação à pontuação do tópico original (sem nenhuma palavra intrusa) que esta representa. Ainda, estão destacados na cor verde os valores da métrica menos sensível, e na cor vermelha os valores da métrica mais sensível.

A sensibilidade S da métrica M, para os tópicos de tamanho 5 do assunto $A \in$

[8] https://matplotlib.org/index.html

$(Sports, Politics, Religion, Music, Christmas)$, é calculada pela seguinte fórmula proposta:

$$S_M(A_5) = \frac{1}{2} \cdot \sum_{i=1}^{2} Abs(M(A_{5-0}) - M(A_{5-i}))$$

A sensibilidade S da métrica M, para os tópicos de tamanho 10 do assunto $A \in (Sports, Politics, Religion, Music, Christmas)$, é calculada pela seguinte fórmula proposta:

$$S_M(A_{10}) = \frac{1}{4} \cdot \sum_{i=1}^{4} Abs(M(A_{10-0}) - M(A_{10-i}))$$

5.2.3 Análise: Tópicos sobre *Sports*

O gráfico da Figura 5.1 apresenta as pontuações obtidas pelos tópicos do assunto *Sports* em cada uma das métricas. A Tabela 5.2 apresenta a sensibilidade calculada de cada métrica para os tópicos sobre *Sports*.

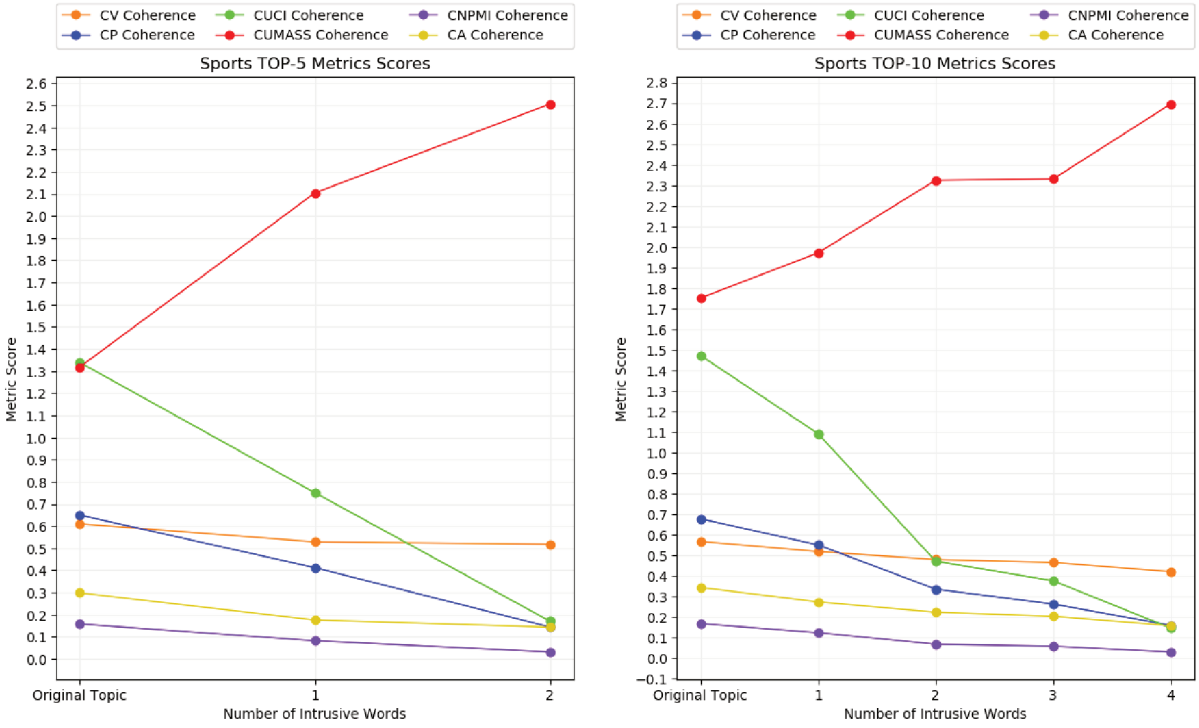

Figura 5.1 – Gráfico das pontuações por métrica do assunto *Sports*.

Mediante análise dos gráficos e da tabela de sensibilidade, é possível observar que para o assunto *Sports*:

- para os tópicos de tamanho 5 a métrica mais sensível foi C_{UMASS};
- para os tópicos de tamanho 10 a métrica mais sensível foi C_{UCI};

Tópicos	C_V	C_P	C_{UCI}	C_{UMASS}	C_{NPMI}	C_A
$Sports_5$	0,08663	0,37127	0,87817	0,98679	0,10102	0,13831
	14,20%	57,16%	65,60%	74,87%	63,83%	46,36%
$Sports_{10}$	0,09656	0,35018	0,94984	0,57832	0,09774	0,12893
	16,99%	51,63%	64,53%	33,02%	57,93%	37,43%

Tabela 5.2 – Sensibilidade das métricas para os tópicos sobre *Sports*.

- para os tópicos de tamanho 5 a métrica menos sensível foi C_V;
- para os tópicos de tamanho 10 a métrica menos sensível foi C_V;
- para todas as métricas, como esperado, quanto mais palavras intrusas, pior a pontuação.

5.2.4 Análise: Tópicos sobre *Politics*

O gráfico da Figura 5.2 apresenta as pontuações obtidas pelos tópicos do assunto *Politics* em cada uma das métricas. A Tabela 5.3 apresenta a sensibilidade calculada de cada métrica para os tópicos sobre *Politics*.

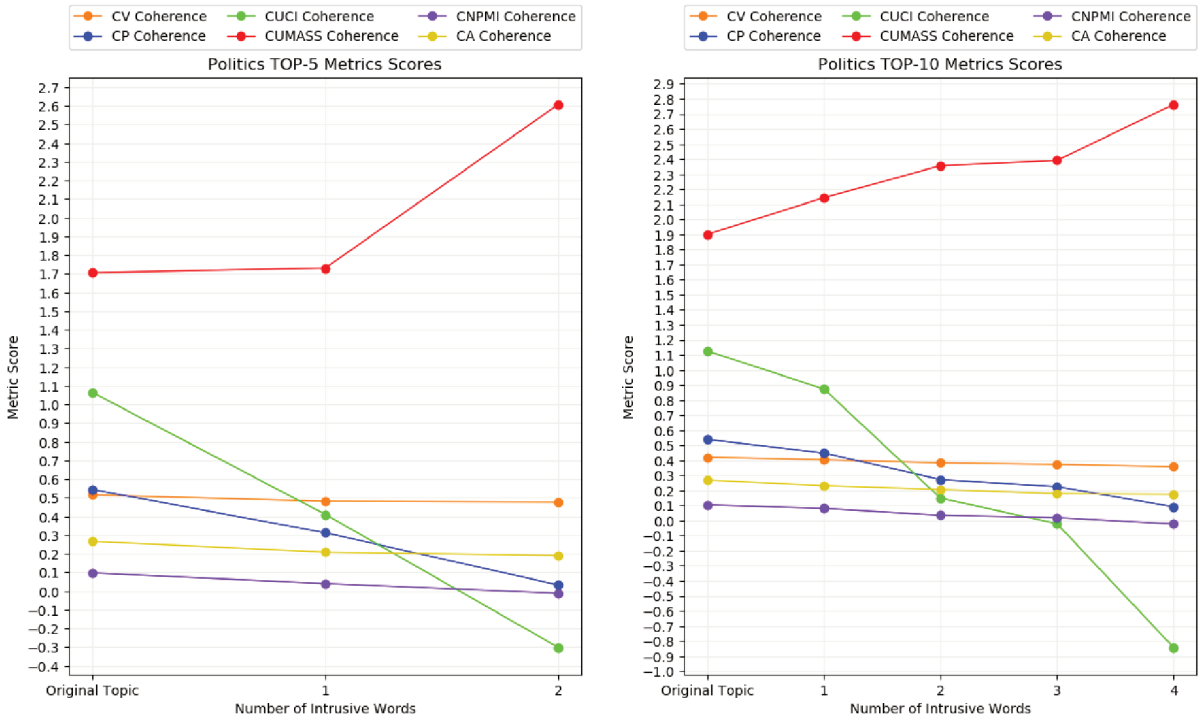

Figura 5.2 – Gráfico das pontuações por métrica do assunto *Politics*.

Mediante análise dos gráficos e da tabela de sensibilidade, é possível observar que para o assunto *Politics*:

- para os tópicos de tamanho 5 a métrica mais sensível foi C_{UCI};

Tópicos	C_V	C_P	C_{UCI}	C_{UMASS}	C_{NPMI}	C_A
$Politics_5$	0,03689	0,37153	1,01005	0,46291	0,08238	0,06624
	7,14%	68,19%	94,83%	27,13%	84,38%	24,83%
$Politics_{10}$	0,04172	0,27936	1,08563	0,51279	0,07535	0,07076
	9,91%	51,87%	96,47%	26,98%	72,75%	26,31%

Tabela 5.3 – Sensibilidade das métricas para os tópicos sobre *Politics*.

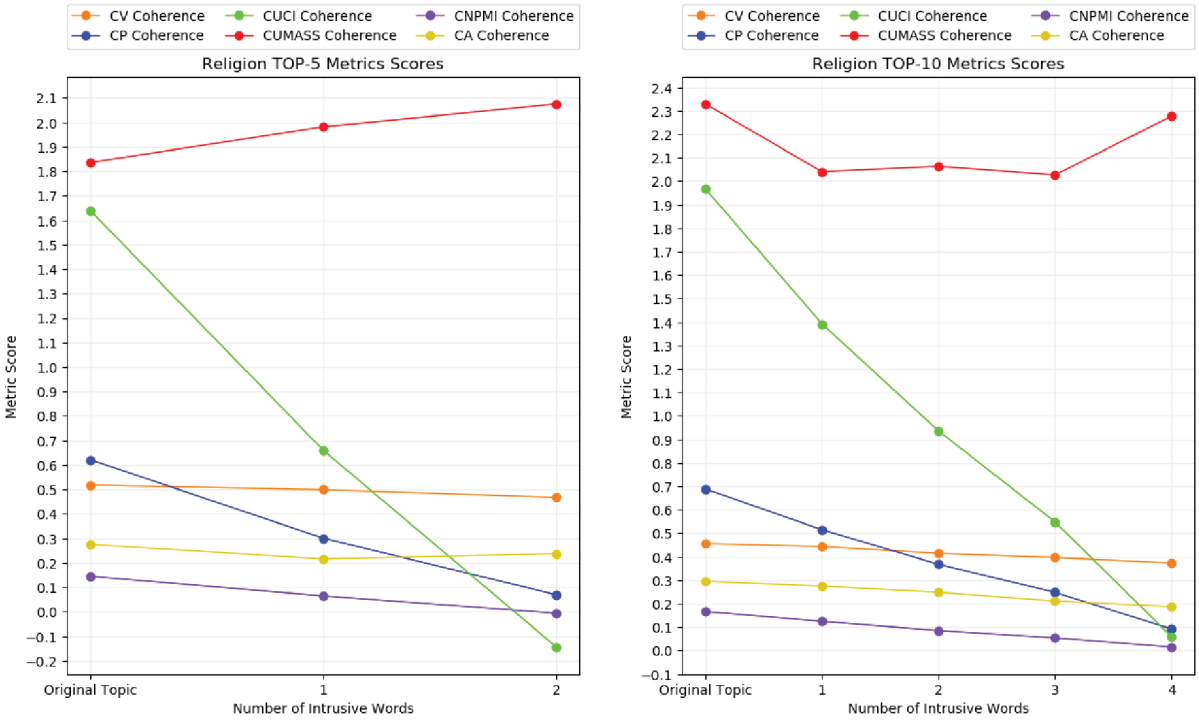

Figura 5.3 – Gráfico das pontuações por métrica do assunto *Religion*.

- para os tópicos de tamanho 10 a métrica mais sensível foi C_{UCI};
- para os tópicos de tamanho 5 a métrica menos sensível foi C_V;
- para os tópicos de tamanho 10 a métrica menos sensível foi C_V;
- para todas as métricas, quanto mais palavras intrusas, pior a pontuação.

5.2.5 Análise: Tópicos sobre *Religion*

O gráfico da Figura 5.3 apresenta as pontuações obtidas pelos tópicos do assunto *Religion* em cada uma das métricas. A Tabela 5.4 apresenta a sensibilidade calculada de cada métrica para os tópicos sobre *Religion*.

Mediante análise dos gráficos e da tabela de sensibilidade, é possível observar que para o assunto *Religion*:

Tópicos	C_V	C_P	C_{UCI}	C_{UMASS}	C_{NPMI}	C_A
$Religion_5$	0,03411	0,43638	1,38186	0,19251	0,11519	0,04855
	6,59%	70,23%	84,27%	10,49%	79,65%	17,66%
$Religion_{10}$	0,04897	0,38342	1,23351	0,22715	0,09648	0,06419
	10,75%	55,66%	62,70%	9,75%	58,06%	21,80%

Tabela 5.4 – Sensibilidade das métricas para os tópicos sobre *Religion*.

- para os tópicos de tamanho 5 a métrica mais sensível foi C_{UCI};
- para os tópicos de tamanho 10 a métrica mais sensível foi C_{UCI};
- para os tópicos de tamanho 5 a métrica menos sensível foi C_V;
- para os tópicos de tamanho 10 a métrica menos sensível foi C_{UMASS};
- para os tópicos de tamanho 5, a métrica C_A teve um comportamento de queda e ascensão;
- para os tópicos de tamanho 10, a métrica C_{UMASS} teve um comportamento de queda e ascensão.

5.2.6 Análise: Tópicos sobre *Music*

O gráfico da Figura 5.4 apresenta as pontuações obtidas pelos tópicos do assunto *Music* em cada uma das métricas. A Tabela 5.5 apresenta a sensibilidade calculada de cada métrica para os tópicos sobre *Music*.

Tópicos	C_V	C_P	C_{UCI}	C_{UMASS}	C_{NPMI}	C_A
$Music_5$	0,03583	0,34079	0,81806	0,40983	0,07357	0,07245
	6,94%	62,45%	70,24%	17,67%	64,57%	36,30%
$Music_{10}$	0,05870	0,33259	0,86543	0,20540	0,06848	0,05838
	12,81%	52,21%	61,57%	8,54%	54,90%	27,25%

Tabela 5.5 – Sensibilidade das métricas para os tópicos sobre *Music*.

Mediante análise dos gráficos e da tabela de sensibilidade, é possível observar que para o assunto *Music*:

- para os tópicos de tamanho 5 a métrica mais sensível foi C_{UCI};
- para os tópicos de tamanho 10 a métrica mais sensível foi C_{UCI};
- para os tópicos de tamanho 5 a métrica menos sensível foi C_V;

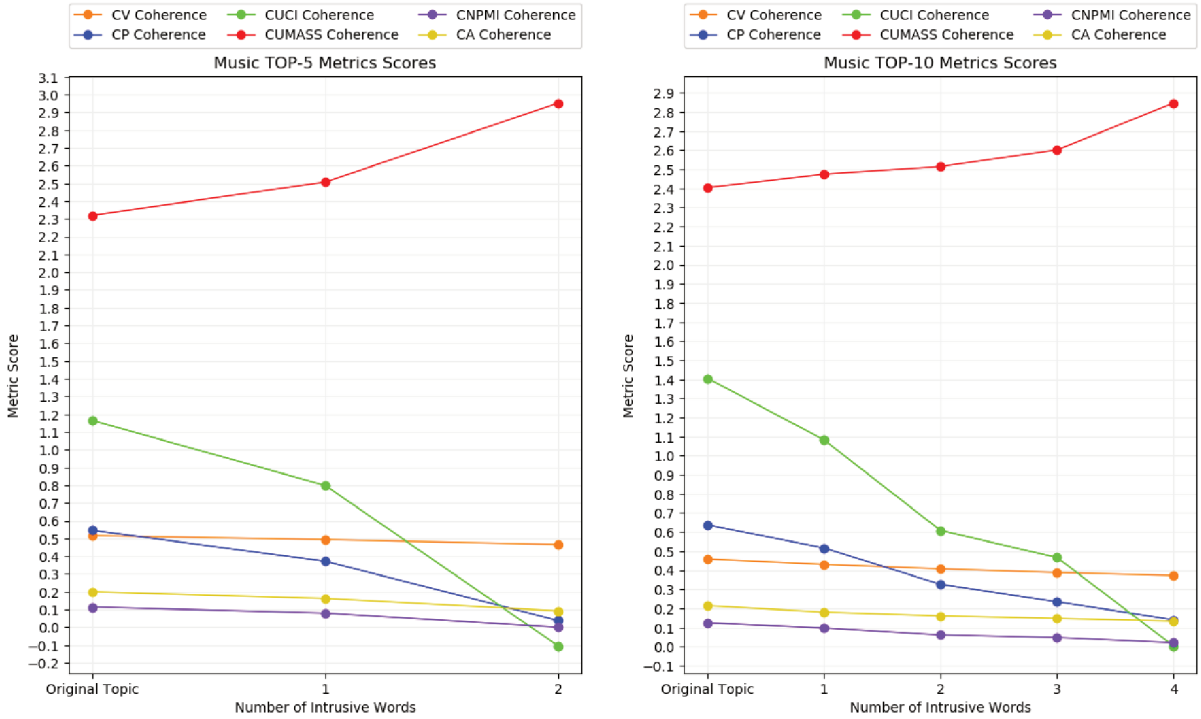

Figura 5.4 – Gráfico das pontuações por métrica do assunto *Music*.

- para os tópicos de tamanho 10 a métrica menos sensível foi C_{UMASS};
- para todas as métricas, quanto mais palavras intrusas, pior a pontuação.

5.2.7 Análise: Tópicos sobre *Christmas*

O gráfico da Figura 5.5 apresenta as pontuações obtidas pelos tópicos do assunto *Christmas* em cada uma das métricas. A Tabela 5.6 apresenta a sensibilidade calculada de cada métrica para os tópicos sobre *Christmas*.

Tópicos	C_V	C_P	C_{UCI}	C_{UMASS}	C_{NPMI}	C_A
$Christmas_5$	0,00230	0,21027	1,21116	0,46374	0,06707	0,00626
	0,49%	57,16%	120,19%	22,83%	83,39%	3,89%
$Christmas_{10}$	0,01850	0,20228	0,94847	0,14813	0,05813	0,02486
	5,25%	60,36%	113,84%	5,25%	89,54%	15,36%

Tabela 5.6 – Sensibilidade das métricas para os tópicos sobre *Christmas*.

Mediante análise dos gráficos e da tabela de sensibilidade, é possível observar que para o assunto *Christmas*:

- para os tópicos de tamanho 5 a métrica mais sensível foi C_{UCI};
- para os tópicos de tamanho 10 a métrica mais sensível foi C_{UCI};

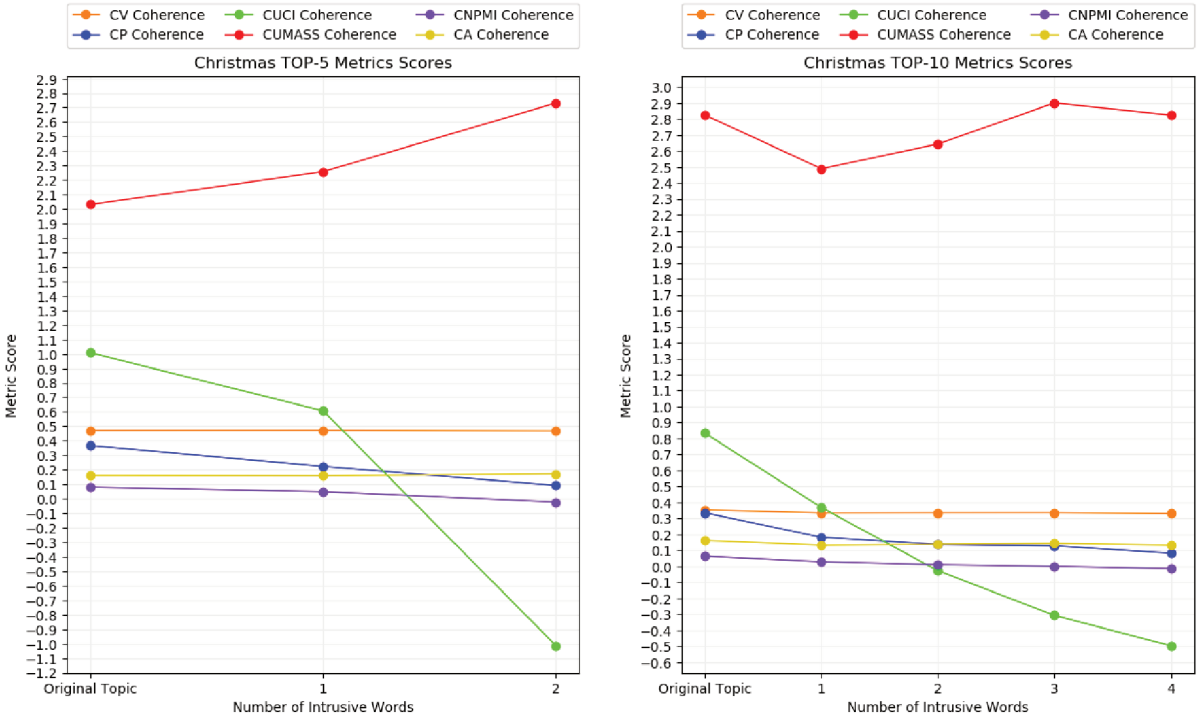

Figura 5.5 – Gráfico das pontuações por métrica do assunto *Christmas*.

- para os tópicos de tamanho 5 a métrica menos sensível foi C_V;
- para os tópicos de tamanho 10 houve um empate técnico, e as métricas menos sensíveis foram C_V e C_{UMASS};
- para os tópicos de tamanho 5, as métricas C_V e C_A tiveram um comportamento de queda e ascensão;
- para os tópicos de tamanho 10, as métricas C_V, C_{UMASS} e C_A tiveram um comportamento de queda e ascensão;

5.2.8 Análise Geral da Sensibilidade das Métricas

A partir das análises realizadas individualmente para cada assunto, percebeu-se uma tendência no comportamento das métricas. Desta forma, uma análise de sensibilidade geral foi realizada sobre as médias das pontuações (Tabela 5.7) de todos os assuntos. Um novo gráfico e uma nova tabela de sensibilidade foram geradas a partir desses novos valores. A Tabela 5.8 e os gráficos da Figura 5.6 apresentam esses valores.

Analisando os gráficos e a tabela de sensibilidade das métricas é possível observar que, em um panorama geral, a métrica que se mostrou mais sensível, tanto para os tópicos de ta-

Tópicos	C_V	C_P	C_{UCI}	C_{UMASS}	C_{NPMI}	C_A
$Formato_{5-0}$	0,52622	0,54586	1,24317	-1,84226	0,11897	0,24014
$Formato_{5-1}$	0,49497	0,32417	0,64551	-2,11636	0,06350	0,18450
$Formato_{5-2}$	0,47944	0,07546	-0,27888	-2,57447	-0,00125	0,16723
$Formato_{10-0}$	0,45116	0,57555	1,36067	-2,24190	0,12563	0,25679
$Formato_{10-1}$	0,42632	0,44276	0,96168	-2,22363	0,09124	0,21892
$Formato_{10-2}$	0,40410	0,28777	0,42771	-2,38024	0,05222	0,19578
$Formato_{10-3}$	0,39175	0,22007	0,21295	-2,44976	0,03606	0,17716
$Formato_{10-4}$	0,37093	0,11335	-0,22596	-2,68117	0,00605	0,15760

Tabela 5.7 – Média da pontuação das métricas por formato de tópico.

manho 5 quanto tamanho 10, foi a C_{UCI}. É possível observar também que a métrica que se mostrou menos sensível para os tópicos de tamanho 5 foi a C_V, e para os tópicos de tamanho 10 foi a C_{UMASS}. Nota-se também que a métrica C_{UMASS} apresentou um comportamento de queda e ascensão devido a este mesmo comportamento apresentado nos tópicos dos assuntos *Religion* e *Christmas*. Ademais, é possível afirmar que de modo geral as métricas refletiram o comportamento esperado mediante a avaliação dos tópicos pelos pesquisados no processo de validação dos mesmos.

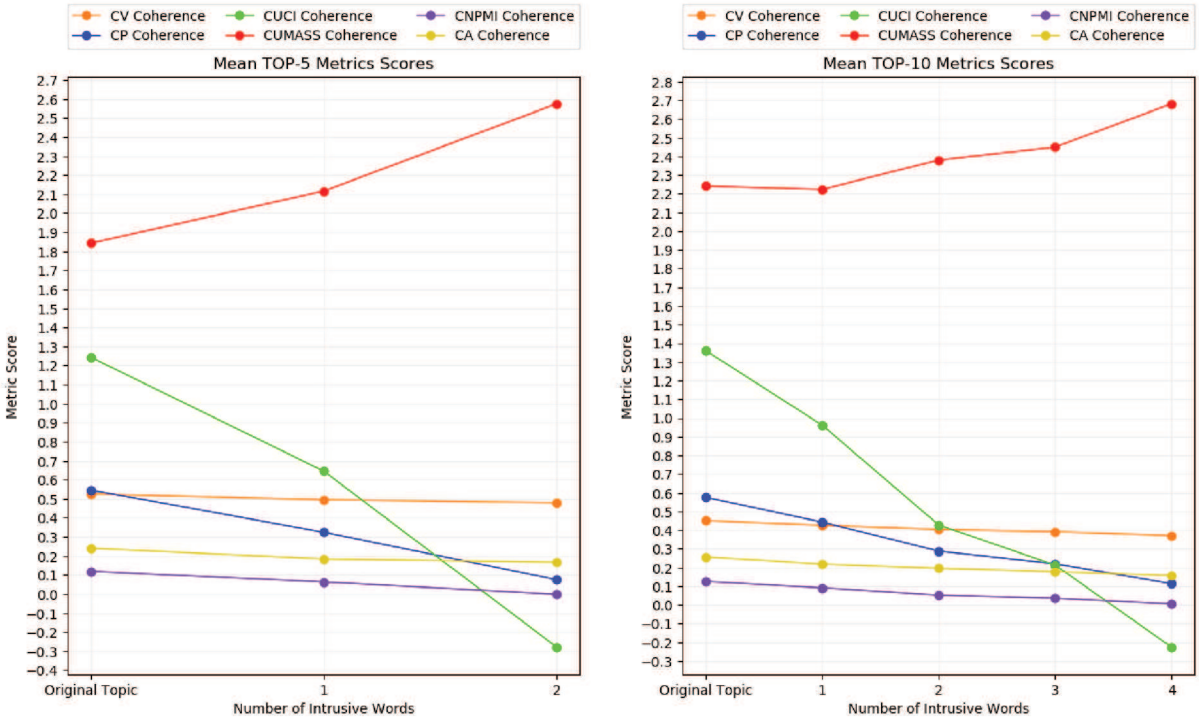

Figura 5.6 – Gráfico das médias das pontuações das métrica por formato de tópico.

Tópicos	C_V	C_P	C_{UCI}	C_{UMASS}	C_{NPMI}	C_A
$Tamanho_5$	0,03901	0,34605	1,05986	0,50316	0,08784	0,06428
	7,41%	63,39%	85,25%	27,31%	73,84%	26,77%
$Tamanho_{10}$	0,05289	0,30957	1,01658	0,20093	0,07924	0,06942
	11,72%	53,79%	74,71%	8,96%	63,07%	27,03%

Tabela 5.8 – Sensibilidade das métricas para a média das pontuações.

5.3 Considerações Finais

Baseando-se nos resultados obtidos e suas análises, constata-se que os comportamentos esperados e atestados pela aplicação dos formulários de validação dos tópicos foi refletido pela aplicação das métricas. Embora tenham sido verificadas algumas anomalias nesse comportamento, de modo geral é possível afirmar que quanto mais palavras intrusas, menor a pontuação do tópico, ou seja, as métricas cumprem o papel de qualificar os tópicos segundo a relação de suas palavras quanto ao contexto.

Os resultados das análises, tanto individuais quanto geral, para os tópicos criados e a métrica de sensibilidade definida, se mostram conclusivos de que: tanto para os tópicos de tamanho 5 quanto de tamanho 10, a métrica mais sensível foi a C_{UCI}; para os tópicos de tamanho 5 a métrica menos sensível foi a C_V, e para os tópicos de tamanho 10 a métrica menos sensível foi a C_{UMASS}. É notório que para os tópicos de tamanho 10 a métrica C_V também se mostrou pouco sensível, e que se não fossem os comportamentos inesperados da métrica C_{UMASS}, esta seria uma forte candidata a métrica mais sensível. Conclui-se também que as duas métricas mais sensíveis foram as que utilizam as medidas de similaridade entre palavras *PMI* e *NPMI*.

5.3.1 Ameaças à Validade

O objetivo deste trabalho é avaliar a sensibilidade das métricas propostas por RÖDER; BOTH; HINNEBURG (2015), e para isso um conjunto de tópicos foi criado e validado através de formulários de pesquisa. Desta forma, a maior ameaça à validade dos resultados obtidos passa por este processo de validação, que está diretamente relacionado as respostas dos pesquisados. Para minimizar esta ameaça, os formulários foram aplicados de forma controlada, e os pesquisados foram selecionados segundo o seguinte perfil de conhecimento: possuir ou estar cursando o ensino superior e ter um nível de conhecimento da língua inglesa de razoável para melhor.

Outra ameaça diz respeito ao número de respostas obtidas quando aplicados os formu-

lários. Porém, acredita-se que pelo fato de que todos os pesquisados foram escolhidos pontualmente, e as respostas foram objetivas e não apresentaram muitas divergências, o número de 7 pessoas por formulário seria suficiente.

Por fim, acredita-se que mediante à minimização dessas ameaças, e embora alguns dos processos possam ser melhorados, os resultados obtidos por este trabalho são conclusivos e relevantes.

6 CONCLUSÃO

Neste trabalho de conclusão de curso, foi realizada uma avaliação de sensibilidade das métricas de avaliação de tópicos propostas em RÖDER; BOTH; HINNEBURG (2015).

Com a crescente tendência de armazenar todo o conhecimento coletivo produzido de forma digital, a tarefa de encontrar todo esse conteúdo tem se tornado cada vez mais difícil (BLEI, 2012). Daí a necessidade de se criar novas ferramentas que ajudem a organizar essa vasta quantidade de dados. A extração de tópicos é uma abordagem que pode ser utilizada para resolver esse problema. Porém, a natureza não-supervisionada desses algoritmos torna difícil a avaliação da corretude interpretativa dos tópicos gerados, tornando dificultosa a tarefa de avaliar a qualidade dos mesmos.

Segundo BLEI (2012), avaliar a qualidade de um tópico manualmente não é viável, uma vez que não temos o poder humano necessário para ler e classificar vastas quantidades de informações. O uso de ferramentas computacionais que implementam métricas para avaliar a qualidade de um tópico é uma das alternativas para tal. Porém, diferentes métricas podem resultar em diferentes resultados, dependendo da forma com que foram implementadas.

No contexto deste trabalho, foram escolhidas 6 métricas para avaliação, sendo elas: C_{UCI} e C_{NPMI}, que são métricas que utilizam as medidas de similaridade *PMI* e *NPMI*, a métrica C_{UMASS}, onde a ocorrência de cada palavra dentre as *N-top words* de um tópico deve ser suportada pelas palavras que a precedem (com maior probabilidade), as métricas C_A e C_V, que utilizam a similaridade através do cosseno entre os vetores gerados pela similaridade dos pares de palavras de um tópico, e a métrica C_P, que calcula a coerência de um tópico através da medida de confirmação de coerência de *Fitelson*.

Para viabilizar a aplicação e avaliação dessas métricas, se fez necessária a criação de um conjunto de tópicos, a partir de um processo que foi dividido nas seguintes etapas: definição de cinco assuntos para criação dos tópicos; escolha das palavras que compõem cada tópico através de um algoritmo de contagem de ocorrência de palavras; criação de dois tópicos (um de tamanho 5 e outro de tamanho 10) a partir das palavras escolhidas; a deturpação dos tópicos através de uma técnica de inserção de palavras intrusas; e por fim a validação dos mesmos por meio da aplicação de formulários de avaliação.

Após a validação dos 40 tópicos criados, os mesmos foram submetidos à aplicação de cada uma das métricas através da plataforma online (versão DEMO) do *Palmetto*. Os resulta-

dos obtidos foram apresentados em forma de tabelas e gráficos. A sensibilidade das métricas foi avaliada a partir de uma métrica de sensibilidade que mede a variação média da pontuação dos tópicos quando palavras intrusas são inseridas nos mesmos. De modo geral, as métricas se comportaram conforme o esperado, onde, apesar de algumas poucas anomalias neste comportamento, quanto mais palavras intrusas nos tópicos, piores as pontuações obtidas pelos mesmos.

Através da análise dos gráficos gerados e da aplicação da métrica de sensibilidade, constatou-se que a métrica C_{UCI} foi a mais sensível, tanto para os tópicos de tamanho 5 quanto para os tópicos de tamanho 10. Ainda, as métricas que se mostraram menos sensíveis foram a C_V e a C_{UMASS}, para os tópicos de tamanho 5 e tamanho 10, respectivamente. A métrica C_V também se mostrou pouco sensível para os tópicos de tamanho 10. Concluiu-se também que as duas métricas mais sensíveis (C_{UCI} e C_{NPMI}), para ambos os tamanhos de tópicos, foram as que utilizam as medidas de similaridade de palavras *PMI* e *NPMI*, respectivamente.

6.1 Trabalhos Futuros

Para abranger as análises realizadas neste trabalho, algumas alternativas podem ser utilizadas durante todo o processo. Uma delas é a implementação das métricas avaliadas em alternativa à ferramenta *Palmetto*, de forma que se tenha um maior controle sobre algumas variáveis como o tamanho das janelas. Outra possibilidade de aprofundar as análises é a utilização de uma outra base de teste que não seja a *Wikipedia*, como por exemplo, utilizar a própria base de dados de onde as palavras que compõem os tópicos foram retiradas. Ainda, é possível a utilização de um conjunto de tópicos maior, e até mesmo de tópicos com mais palavras e diferentes combinações de palavras intrusas.

REFERÊNCIAS

BLEI, D. M. Probabilistic topic models. **Communications of the ACM**, [S.l.], v.55, n.4, p.77–84, 2012.

CHANG, J. et al. Reading tea leaves: how humans interpret topic models. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS. **Anais...** [S.l.: s.n.], 2009. p.288–296.

FITELSON, B. A probabilistic theory of coherence. **Analysis**, [S.l.], v.63, n.279, p.194–199, 2003.

MIMNO, D. et al. Optimizing semantic coherence in topic models. In: OF THE CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING. **Proceedings...** [S.l.: s.n.], 2011. p.262–272.

RÖDER, M.; BOTH, A.; HINNEBURG, A. Exploring the Space of Topic Coherence Measures. In: ACM INTERNATIONAL CONFERENCE ON WEB SEARCH AND DATA MINING. **Proceedings...** [S.l.: s.n.], 2015. p.399–408.

STEYVERS, M.; GRIFFITHS, T. Probabilistic topic models. **Handbook of latent semantic analysis**, [S.l.], v.427, n.7, p.424–440, 2007.